Escola d'Enginyeria de Telecomunicació i Aeroespacial de Castelldefels
UNIVERSITAT POLITÈCNICA DE CATALUNYA

PROYECTO FINAL DE CARRERA

TÍTULO DEL PFC: Diseño de una solución de datacenter en alta disponibilidad. Ventajas e inconvenientes de utilizar virtualización de red.

TITULACIÓN: Ingeniería de Telecomunicación

AUTOR: Roberto Galera Marí

DIRECTOR: David Remondo Bueno

FECHA: 1 de febrero de 2017

TÍTULO: Diseño de una solución de datacenter en alta disponibilidad. Ventajas e inconvenientes de utilizar virtualización de red.

AUTOR: Roberto Galera Marí

DIRECTOR: David Remondo Bueno

FECHA: 1 de febrero del 2017

Resumen

En este proyecto, en base a los requerimientos de un cliente para un entorno real, se ha visto los pasos necesarios y los retos que se encuentran al diseñar una solución de datacenter en alta disponibilidad distribuida entre distintas localizaciones.

En primer lugar, se ha optado por realizar este diseño de manera tradicional. Esto es, utilizando equipos físicos para realizar las distintas funciones de red necesarias en la arquitectura. Se ha hecho el diseño inicial y, una vez escogidos los equipos y teniendo en cuenta las tecnologías que utilizan, se ha propuesto la topología física.

De manera alternativa a este diseño, se propone la implementación de la solución para el cliente utilizando una arquitectura virtual basada en SDDC (Software Defined DataCenters) y SDN (Software Defined Networks), tecnologías que actualmente se están empezando a utilizar para substituir el uso de arquitecturas tradicionales en datacenters. En este caso se ha utilizado la arquitectura NSX de VMWare para hacer esta virtualización. Para el diseño de la solución se han utilizado la tecnología de overlay VXLANs, servicios de red distribuidos y servicios de red desplegados sobre appliances virtuales que utiliza NSX para ofrecer, desde el punto de vista funcional, los mismos servicios.

El objetivo es comprobar si es posible satisfacer todos los requerimientos de un cliente real utilizando virtualización de red, y ver qué ventajas e inconvenientes presenta el uso de esta tecnología, en relación a la arquitectura tradicional propuesta.

Title: Design of a datacentre solution with high availability. Advantages and disadvantages of using network virtualization.

Author: Roberto Galera Marí

Director: David Remondo Bueno

Date: February 1st, 2017

Overview

In this project, based on the requirements of a client for a real environment, we have seen the necessary steps and challenges that are encountered when designing a high availability datacentre solution distributed among different locations.

In the first place, this design has been done in a traditional way. That is, using physical equipment to perform the various network functions required in the architecture. The initial design has been made and, once the devices have been chosen and considering the technologies they use, the physical topology has been proposed.

As an alternative to this design, it is proposed to implement the solution for the client using a virtual architecture based on Software Defined DataCentres (SDDC) and SDN (Software Defined Networks), technologies that are currently being used to replace the use of traditional architectures in datacentres. In this case VMWare NSX architecture has been used to do this virtualization. To design this solution, VXLANs overlay technology, distributed network services and network services deployed over virtual appliances used by NSX have been used to offer, from the functional point of view, the same services.

The objective is to verify if it is possible to satisfy all the requirements of a real client using network virtualization, and to see what advantages and disadvantages the use of this technology presents, in relation to the traditional architecture proposed.

INDICE

INTRODUCCIÓN

Desde hace años se ha extendido entre las empresas el uso de centros de datos (o datacenters) para mantener de manera unificada su infraestructura de servidores. El aumento de la velocidad en las redes de acceso de última milla (como accesos por fibra, cable o 4G en redes móviles) y el aumento exponencial de la capacidad de transporte en las redes troncales, ha permitido que la conexión a los servidores en esta infraestructura se pueda hacer con los mismos valores de velocidad, latencia y retardo que se haría si estuvieran conectados a la red local de cada oficina.

La utilización de infraestructuras en datacenter permite a las empresas evitar los problemas de tener su infraestructura de servidores distribuida a lo largo de sus distintas localizaciones o en su oficina central como son: un costoso espacio inmobiliario para ubicar sus servidores, un alto coste de mantenimiento o la necesidad de personal cualificado en cada localización para la gestión de estos equipos. Además, se pueden despreocupar de otras consideraciones que se deben tomar para asegurar la seguridad y la alta disponibilidad de sus servidores, y que se ofrecen en arquitecturas de datacenter: disponer acometidas eléctricas redundantes, tener sistemas de alimentación ininterrumpida basados en baterías y generadores diésel si falla la red pública, tener mecanismos de refrigeración redundantes, tener equipos para asegurar la seguridad de accesos y prevención de intrusiones, así como contar con sistemas para la protección contra incendios y catástrofes.

El hecho de que las grandes empresas pueden tener muchas sedes repartidas en distintos países y que muchas de estas basen total o parcialmente su negocio en la venta de sus productos o servicios online, hacen que los requerimientos en cuanto a disponibilidad sean muy estrictos. A pesar de que muchos datacenters se construyen asegurando disponibilidades cercanas al 100% del tiempo, un problema imprevisto podría hacer que toda la infraestructura del cliente quedase aislada. Por esta razón, cuando los requerimientos de disponibilidad son muy altos se opta por tener infraestructuras de respaldo en otros datacenters a las que balancear el tráfico en caso de desastre.

El uso creciente de virtualización para los servidores en los datacenters (Software Defined DataCenters, SDDC) ha permitido aprovechar al máximo los recursos desplegando múltiples servidores virtuales sobre cada hipervisor, aprovechando al máximo los recursos, aumentando la eficiencia energética y reduciendo el coste total. Además, la virtualización permite de manera flexible y ágil desplegar nuevos servidores o cambiar sus características de acuerdo a las necesidades de las aplicaciones. Sin embargo, esto no ocurre con las redes que los soportan, ya que estas necesitan una configuración manual. Además, esta configuración es muy dependiente de los protocolos utilizados y la manera en que cada fabricante los implementa. Por esta razón un cambio en las necesidades de las aplicaciones que a nivel de servidor se hace de manera ágil

puede requerir cambios a nivel de red del orden de días o semanas y en algunos casos el rediseño total o parcial de esta red.

En este contexto se está empezando a implantar el uso de redes definidas por software (SDN, Software Defined Networks). Éstas, desacoplando el plano de datos y el de control de la red, permiten tener un fabric (substrato) único sobre el que se programan y despliegan los requerimientos de red que necesitan los servidores y las aplicaciones en cada momento. El controlador está centralizado y permite hacer esta programación de manera automática. En general el uso de SDN permite, teniendo visión global de toda la red, la provisión de servicios de red bajo demanda sobre un entorno multi-fabricante. Esto, en entornos de grandes datacenters por ejemplo permite la programación de las redes desplegadas desde un único punto de acceso utilizando APIs. Además, en entornos de *Cloud*, mediante el uso de orquestadores con acceso a este API y a la de la virtualización de los servidores, permite el despliegue de las infraestructuras de sus clientes de manera automática extremo a extremo.

Desde el punto de vista académico, existen muchos estudios sobre las tecnologías que se utilizan en las redes tradicionales y, actualmente, mucho interés en el estudio de estándares y funcionalidades para redes SDN. A veces, las soluciones teóricas que se estudian no se corresponden al 100% con las implementaciones prácticas de los fabricantes para su utilización en entornos reales.

De acuerdo con las observaciones anteriores, en este proyecto se propone el diseño de una solución de red en base a los requerimientos de un cliente para un entorno real; dicha solución requiere alta disponibilidad y estará distribuida entre dos datacenters para asegurar la disponibilidad del servicio en caso de fallo grave en alguno de ellos. Esta solución la utilizará el cliente para desplegar sus servidores y proporcionar servicios tanto internos para los empleados en sus distintas localizaciones como para sus clientes a través de Internet.

En primer lugar, se propone hacer este diseño de red de manera tradicional. Esto es, utilizando equipos físicos para realizar las distintas funciones de red necesarias en la arquitectura. Se pretende ver cuáles son los pasos necesarios y los retos a la hora de diseñar esta solución. Además, estudiar los equipos que hay en el mercado que más se ajusten a las necesidades del cliente.

De manera alternativa, se propone el diseño de la misma solución utilizando virtualización de red. Esto se hará utilizando la arquitectura NSX de VMWARE para la virtualización de red junto a la virtualización de los servidores (SDN y SDDC). Se pretende ver si esta arquitectura, que separa la red física subyacente de la virtual utilizando tecnología de overlay VXLANs, servicios de red distribuidos y servicios de red utilizando appliances virtuales permite ofrecer de manera funcional los mismos servicios que ofrece la solución tradicional. También tiene como objetivo ver qué ventajas e inconvenientes supone el uso de SDN, y NSX en particular, a la hora de diseñar una solución de estas características.

Partiendo de estos dos diseños de red, se pretende discutir qué ventajas e inconvenientes suponen el uso de una arquitectura tradicional basada en equipos físicos y una arquitectura basada en el uso de virtualización de red a la hora de implementar una solución real de datacenter en base a los requerimientos de un cliente.

Este proyecto se compone de 4 capítulos. En el capítulo 1 se explican los requerimientos de un cliente para la solución de datacenter. En el capítulo 2 se expone en base a estos requerimientos la propuesta de diseño de red tradicional basada en equipos físicos. En el capítulo 3 se describe la propuesta de diseño basada en virtualización de red. Finalmente, en el capítulo 4 se discuten las ventajas e inconvenientes de las dos propuestas a la hora de implementar la solución del cliente.
.

CAPÍTULO 1. REQUERIMIENTOS

1.1. Requerimientos del cliente

El punto de partida para este proyecto son los requerimientos presentados por un cliente real para desplegar una infraestructura de red en un centro de datos (datacenter o DC a partir de ahora) con el fin de proporcionar una serie de servicios tanto a sus clientes como a sus distintos empleados.

El cliente planea proporcionar una solución de ERP (Enterprise Resource Planning) para 30000 empleados y colaboradores que se conectarán de forma remota tanto a través de Internet como desde sus oficinas utilizando enlaces dedicados. También pretende proporcionar un servicio de Sharepoint con el que crearán sitios web que se utilizarán como lugar seguro donde almacenar, organizar y compartir información. Alguno de estos servicios será accesible directamente o a través de una plataforma Citrix dónde los usuarios se conectarán y desde allí accederán de manera segura a los servicios internos. Además, se utilizará la solución de DC para publicar hacia Internet los sitios públicos que usarán sus potenciales clientes para obtener información, así como los sitios con zonas privadas para que los clientes puedan realizar sus gestiones con la compañía.

Por el carácter crítico de estos servicios y el impacto directo sobre el negocio, se exige una disponibilidad del 99.99%. Para ello se requiere una plataforma con las siguientes características:

- Que la plataforma esté distribuida en dos DCs tier 3 o superior y que cada uno de ellos pueda ofrecer el servicio completo en caso de desastre.
- Conectividad con un tercer DC en el que se puedan almacenar copias de seguridad (backups).
- Los dos DCs principales deben estar interconectados por 2 enlaces a 1 Gbps para la replicación de datos y tener cada uno una salida a Internet a 100 Mbps.
- Además, cada DC tiene que tener un enlace a 100 Mbps con el DC de backup y otro a 100 Mbps contra las oficinas centrales del cliente.
- El cliente gestionará una serie de *Blade Chassis servers* en cada localización (en concreto 3 en cada una de ellas) donde desplegará sus servidores.
- El proveedor de servicios tiene que proveer una infraestructura de red en cada DC de acuerdo a los servicios de red explicados a continuación.

1.1.1. Servicios de balanceo de carga

El servicio tiene que proveer balanceo de carga de manera dinámica, transparente y escalable, distribuyendo y balanceando el tráfico en una granja de servidores Web localizados en la misma localización o en múltiples localizaciones distribuidas geográficamente. El servicio tiene que dar la posibilidad al cliente de dar un único link (nombre de dominio) a los usuarios para el servicio que los redirija de manera transparente al siguiente equipo disponible, dando la apariencia de un único servidor. El servicio debe también redirigir el servicio desde servidores que estén offline a servidores que estén disponibles. Esto incluye la conmutación automática y transparente en caso de fallo sin tener en cuenta la localización de los servidores.

Funcionalidades técnicas necesarias:

- Persistencia de sesión extremo a extremo, desde el cliente al servidor de destino, utilizando cookies de sesión.
- Posibilidad de poner los equipos de balanceo de carga (load balancers, LBs) en clúster para evitar puntos únicos de fallo.
- Cada LB tiene que ser capaz de soportar 5000 sesiones concurrentes como mínimo absoluto.
- Monitores (health checks) en capa 7 que por ejemplo permitan sacar un servidor de la tarea de balanceo sin interrupción del servicio.
- Aceleración SSL (Secure Socket Layer).
- Balanceo basado en políticas como por ejemplo selección del que tenga menos conexiones pendientes, round robin, o round robin con pesos.
- Capacidad de crear tanto clústeres como grupos de balanceo simples.
- Capaz de facilitar la terminación SSL con certificados wildcard (comodín).
- Capaz de realizar la autenticación de sesiones de Citrix (p.ej. F5 APM, Access Policy Manager).

Por lo que respecta al balanceo geográfico se debe hacer utilizando un sistema basado en resolución del Domain Name System (DNS). El cliente delegará la resolución de los dominios manejados por la solución a estos equipos. El proveedor se encargará de monitorizar y mantener el correcto funcionamiento de los servidores DNS y las tablas/bases de datos de resolución de nombres.

1.1.2. Servicios de Seguridad

Se requiere la gestión de firewalls que se encarguen de permitir, limitando o previniendo accesos externos. El propósito de esta limitación es asegurar los sistemas frente ataques o accesos no autorizados. Se requiere un especial énfasis en la seguridad de los servidores internos en la zona protegida de la infraestructura.

Además, de forma opcional, se requiere el estudio y viabilidad de una solución de inspección profunda del tráfico en busca de vulnerabilidades y ataques (IDS, Intrussión Detection System) para el tráfico que viene del exterior.

1.1.3. Red de área local (LAN)

Este servicio proporciona la gestión de los equipos LAN de Capa 2 que se utilizarán para la conectividad entre los servidores instalados en el DC y otros equipos de red. Típicamente, para cubrir esta necesidad en cada site se utilizan conmutadores (switches) tanto L2 como L2/L3. Se requiere monitorización, notificación, detección de congestión, así como la administración y configuración de los equipos.

Las características de los equipos dependerán del tier en el que se apliquen. Los requerimientos más importantes son:

- Conectividad al menos a 1 Gbps en el tier perimetral.
- Conectividad a 10 Gbps para el tier protegido (producción): 16 puertos x 3 blade chassis.
- Alta disponibilidad a nivel de equipo. (N+1) en cada DC.

Además, la infraestructura se debe permitir la comunicación entre una serie de redes en los dos DCs para asegurar tareas de replicación.

1.1.4. Servicios de conectividad

Incluye los servicios de Conectividad entre los DCs donde se montará la infraestructura y hacia las localizaciones remotas. Como se ha comentado anteriormente:

- 2x1 Gbps Enlace entre las dos localizaciones para la replicación de datos.
- 1x100 Mbps. Salida a Internet en cada localización.
- 1x100 Mbps. Enlace con el site donde se almacenan las copias de seguridad en cada localización.
- 1x100 Mbps. Enlace con las oficinas del cliente en cada localización. Desde aquí habrá acceso al resto de oficinas del cliente.

CAPÍTULO 2. PROPUESTA DE DISEÑO TRADICIONAL BASADA EN EQUIPOS FÍSICOS

2.1. Propuesta de diseño

En esta sección se presenta una propuesta de diseño tradicional, basada en equipos físicos, para la solución del cliente de acuerdo a sus requerimientos. El diseño se puede ver a alto nivel en la figura 2.1.

La propuesta consiste en la implementación de dos plataformas en dos DCs basada en dos zonas: una zona perimetral y una zona protegida. Los dos DCs estarán interconectados para asegurar la replicación de ambas plataformas (algunas redes estarán extendidas en Capa 2 y otras en Capa 3 utilizando una red de tránsito INTER DC) y cada uno de ellos estará conectado a las oficinas del cliente y el site de backup. Además, cada DC dispondrá de una salida a Internet dual a 100 Mbps. Todos los equipos y enlaces serán desplegados con redundancia (1+1) para asegurar la disponibilidad en caso de fallo.

En la zona perimetral se encontrarán los servicios directamente conectados con las redes externas (Internet, oficinas del cliente…) así como la DMZ (*DeMilitarized Zone*) donde se encontrarán los servicios directamente publicados, que estarán balanceados para asegurar la distribución de la carga y la alta disponibilidad en caso de fallo de alguno de los servidores. En esta zona también se encontrarán los balanceadores globales que se encargarán de distribuir la carga entre ambos DCs o a uno de ellos en caso de desastre. Para ello en esta zona contará con los siguientes elementos de red:

- Dos firewalls perimetrales en modo activo/pasivo que se encargarán de interconectar y proveer de seguridad a las redes en esta zona. También proporcionarán servicios de VPN (*Virtual Private Network*) LAN-to-LAN para los distintos proveedores del cliente, así como VPNs de acceso remoto para que el cliente se pueda conectar también de forma remota.
- Dos LB locales en modo activo/pasivo: Realizarán el balanceo del tráfico entre los distintos servidores en la DMZ del cliente, así como el balanceo entre los servidores en la zona interna de la plataforma accesibles desde la DMZ.
- Un LB global por datacenter en modo activo/activo que, utilizando DNS, se encargará del balanceo de tráfico entre ambas plataformas. También realizará la conmutación automática en caso de que una de ellas falle.
- Dos switches para asegurar alta disponibilidad: Se encargarán de la interconexión de los elementos de red en la zona protegida. La conectividad se proporcionará a 1 Gbps.

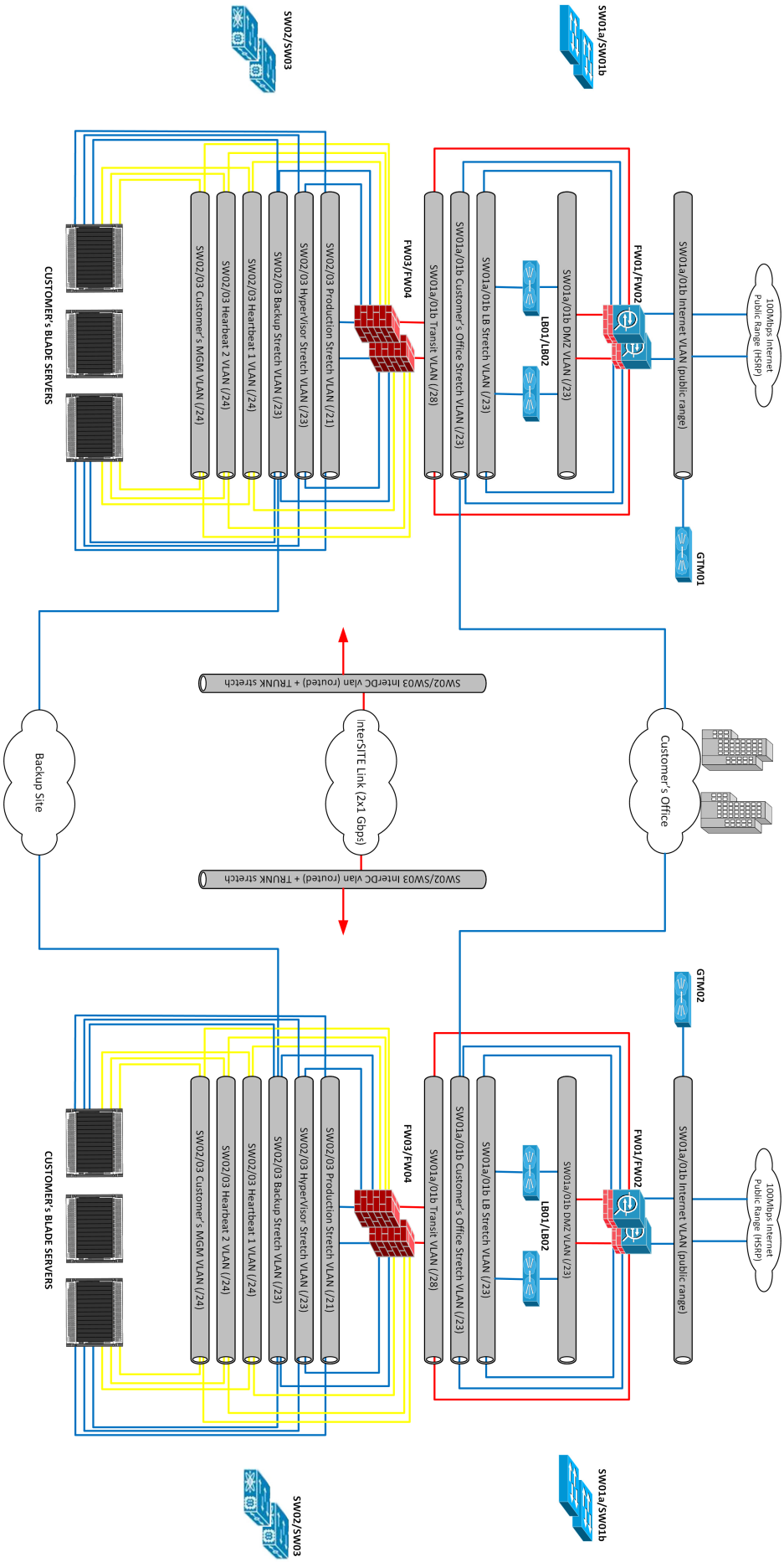

Fig.2.1 Propuesta de diseño.

En la zona protegida se encontrarán los servicios críticos del cliente como los servidores de aplicaciones y las bases de datos que no serán accesibles directamente desde las redes externas. Para esta zona se han utilizado:

- Dos firewalls protegidos en alta disponibilidad en modo activo/pasivo que se encargarán de proveer seguridad entre las redes en esta zona. También representarán un segundo nivel de seguridad para las conexiones que vengan desde las redes externas.
- Dos switches para asegurar alta disponibilidad. Se encargarán de la interconexión de los equipos de esta zona con los blade chassis del cliente que contienen los servidores. Su principal característica es la conectividad a 10 Gbps.

A Nivel 3 ambas zonas estarán interconectadas por una red de tránsito entre los firewalls perimetrales y protegidos y a Nivel 2 la topología de switches utilizará el Spanning-Tree Protocol (STP) para asegurar la redundancia. También se utilizarán otras tecnologías como LACP (Link Agregation Control Protocol) para la agregación de enlaces y VPC (Virtual Port Channels) para la agregación de enlaces entre múltiples dispositivos. Estas tecnologías se utilizarán en los enlaces redundantes entre dispositivos, para proveer enlaces de capacidad agregada y al mismo tiempo superar la limitación que supone utilizar STP donde algunos enlaces se encuentran bloqueados. Todo esto se encuentra explicado en las consideraciones de diseño a continuación.

2.2. Consideraciones de diseño

En esta sección se presentan las consideraciones que se han tenido en cuenta para hacer el diseño de la solución.

2.2.1. Asignación de IPs y VLANs

La asignación de IPs se ha hecho escogiendo subredes en las redes 10.251.0.0/16 y en la 172.30.0.0/16. La primera subred se ha utilizado para las redes usadas por los servidores del cliente para evitar que se solapen en caso que quieran acceder desde sus oficinas. La segunda subred se ha utilizado para redes que, aunque estén presentes en los servidores de cliente, solo tienen sentido local (como las redes de heartbeat o tránsito)

Se han asignado subredes /21, /23 o /24 según las necesidades actuales, pero en todos los casos se ha escogido la primera red dentro de la /21 correspondiente de manera que si fuera necesario en el futuro ampliarla solo habría que cambiar la máscara de red.

Para las redes que necesitan ser desplegadas para comunicar ambos DCs se ha asignado la subred superior al DC01 y la subred inferior al DC02.

Por lo que respecta a las VLANs (Virtual LANs) también se ha tenido en cuenta el posible solapamiento de las VLANs en caso de que las quieran extender

hacia sus oficinas. Para estas VLANs se han utilizado identificadores de VLAN a partir del 900. Para el resto, con sentido local en la infraestructura, se han ido asignando desde la VLAN10.

En la Figura 2.2. se pueden ver las distintas redes y VLANs asignadas. En rojo están las redes conectadas a los firewalls perimetrales, en azul las redes conectadas a los firewalls protegidos y en amarillo la red de tránsito entre ambas zonas. Para ver el uso de cada una de las redes consultar el anexo A.

Nombre VLAN		Stretch (Si/No)	VLAN ID	Redes DC1	Redes DC2
DC1/DC2 Internet		No	499/500	62.97.100.0/24	62.67.101.0/24
DMZ		No	10	172.30.0.0/23	172.30.1.0/23
LB Stretch		Si	60	10.251.240.0/23	10.251.241.0 /23
Customer's Network (Office)		Si	40	192.168.100.0	192.168.100.0
Transit		No	50	172.30.14.0/28	172.30.15.0/28
Production Stretch	Production	Si	900	10.251.248.0 /21	10.251.251.0 /21
	Virtual Hosts	Si	900	10.251.249.0/21	10.251.252.0/21
	Reserved	Si	900	10.251.250.0/21	10.251.253.0/21
	Global Cluster	Si	900	10.251.254.0/21	10.251.254.0/21
	Non	Si	900	10.251.255.0/21	10.251.255.0/21
Heartbeat 1		No	910	172.30.4.0/24	172.30.5.0/24
Heartbeat 2		No	920	172.30.6.0/24	172.30.7.0/24
Backup Stretch		Si	120	10.251.232.0 /23	10.251.233.0 /23
Customer's Management		No	110	10.251.224.0 /24	10.251.225.0 /24
HyperVisor		Si	930	10.251.216.0/23	10.251.217.0/23
InterDC		Si	80	172.30.12.128/26	

Fig. 2.2 Asignación de IPs y VLANs para la solución.

2.2.2. Colocación de los GTMs (Global Traffic Managers)

Se han contemplado dos opciones para la colocación de estos equipos:

a) Colocarlos detrás de los firewalls perimetrales, en la VLAN LB/DMZ.
b) Colocarlos directamente en la VLAN pública de la infraestructura por fuera de los firewalls perimetrales.

La resolución de nombres (DNS) es primordial para el funcionamiento de cualquier servicio publicado en Internet. En el caso de esta infraestructura es especialmente importante porque la resolución DNS se hace de manera dinámica con unos TTLs (Time to Live) bastante bajos para asegurar que la asignación del pool de servidores en cada DC es la óptima teniendo en cuenta el estado actual del servicio. No sirve de nada tener todos los servidores distribuidos en los dos DC disponibles si la resolución DNS no funciona.

El hecho de que los TTLs sean tan bajos hacen que la carga en los servidores DNS autoritativos para los dominios (en este caso los GTMs) sea bastante más alta que en los servidores DNS tradicionales ya que la caché de los clientes caducará con frecuencia, obligando a volver a consultar. En este caso se espera que estos equipos reciban gran cantidad de peticiones (de los 30000 usuarios concurrentes de la empresa para servicios internos y demás usuarios de Internet para paginas públicas).

Se ha decidido poner estos equipos (uno por DC, proporcionando redundancia entre ellos) directamente en la VLAN pública fuera del ámbito de los firewalls. Esto se ha hecho por un lado para que esta gran cantidad de conexiones concurrentes no afecten al rendimiento de los firewalls perimetrales y por otro para asegurar la continuidad del servicio DNS en caso de que caigan por ejemplo estos firewalls. Si esto sucediera, por ejemplo en un DC los pools de servidores dejarían de estar disponibles pero no el GTM. Éste detectaría esta situación y seguiría resolviendo los nombres de dominio, pero con IPs del otro DC, que si estará disponible.

Las ventajas de esta colocación serían que se aumenta la tolerancia a fallos del servicio DNS (al quitar puntos de fallo por delante de estos) y que se evita la necesidad de utilizar firewalls más potentes debido a la gran cantidad de conexiones. El principal inconveniente es que se pierde la protección que proporciona el firewall. Utilizando GTMs F5 esto no sería un problema ya que proporciona mecanismos de seguridad que previenen frente a amenazas como ataques DoS (Denial of Service attacks) contra DNS, envenenamiento de la cache DNS (cache poisoning) o secuestro DNS (DNS hijacking) más allá de lo que podría hacer el propio firewall (filtrando a nivel 3/4) y permite crear políticas que proporcionen un nivel añadido de protección. Algunos de los mecanismos incluidos son:

- Equipos robustos: Los equipos Big IP están certificados como firewalls de red por ICSA labs (ver [1]) y resisten ataques típicos.
- Protección frente a ataques DNS: De manera nativa proporciona validación del protocolo a nivel software, descartando cuando hay alto volumen de paquetes UDP, peticiones DNS, NXDOMAIN floods y paquetes malformados; también se pueden mitigar estos ataques a nivel hardware.
- Control de seguridad: Usando iRules (ver [2]) DNS se pueden crear políticas que ayuden a proteger frente a peticiones que puedan resultar ser una amenaza.
- Filtros de paquetes: Se puede limitar el acceso al servicio basándose en IP de origen, de destino o puerto, como haría el firewall.

El GTM tiene que tener visibilidad sobre las VIPs (Virtual IPs) en los balanceadores locales en cada datacenter para poder evaluar si el servicio está levantado y, por tanto, puede enviar peticiones. En esta arquitectura (con los GTMs en la VLAN pública) esto se hará haciendo las comprobaciones contra las IPs públicas correspondientes a cada VIP. Para asegurar la correcta monitorización será necesario tener abiertos los flujos correspondientes desde la IP del GTM hacia estas VIPs en los firewalls perimetrales.

2.2.3. Uso de stretch VLANs. Comunicación entre DCs en Capa 2.

Se espera una gran cantidad de tráfico dentro del mismo tier dentro de cada DC y entre DCs, especialmente tráfico de replicación entre equipos que desempeñan el mismo rol. Por ejemplo, Servidores Web en la VLAN LB, servidores de aplicaciones y bases de datos en la VLAN de producción. Este tráfico (este-oeste) en la arquitectura se espera que sea de gran volumen. Teniendo en cuenta que se trata del mismo tier y que a priori no hay ningún problema en que las máquinas se vean entre ellas directamente se ha decidido estirar (stretch) algunas de las VLANs donde se espera este tipo de tráfico entre los dos DCs.

Hacer esto también permite aumentar la redundancia y la capacidad de los servicios publicados desde cada DC al poder acceder a los servidores de uno de los DCs desde el otro. Por ejemplo, se podría definir un pool de servidores en el DC2 que contenga servidores en la VLAN LB de ambos DCs (Webfarm DC01 y Webfarm DC02 en la figura 2.3). El balanceador en el DC2 tendría acceso directo al estar en la misma red física. Si fallase la salida a Internet, los firewalls perimetrales, o los balanceadores locales del DC1, se podría seguir usando la capacidad de los servidores en el DC1 (flecha verde) mientras se soluciona el problema.

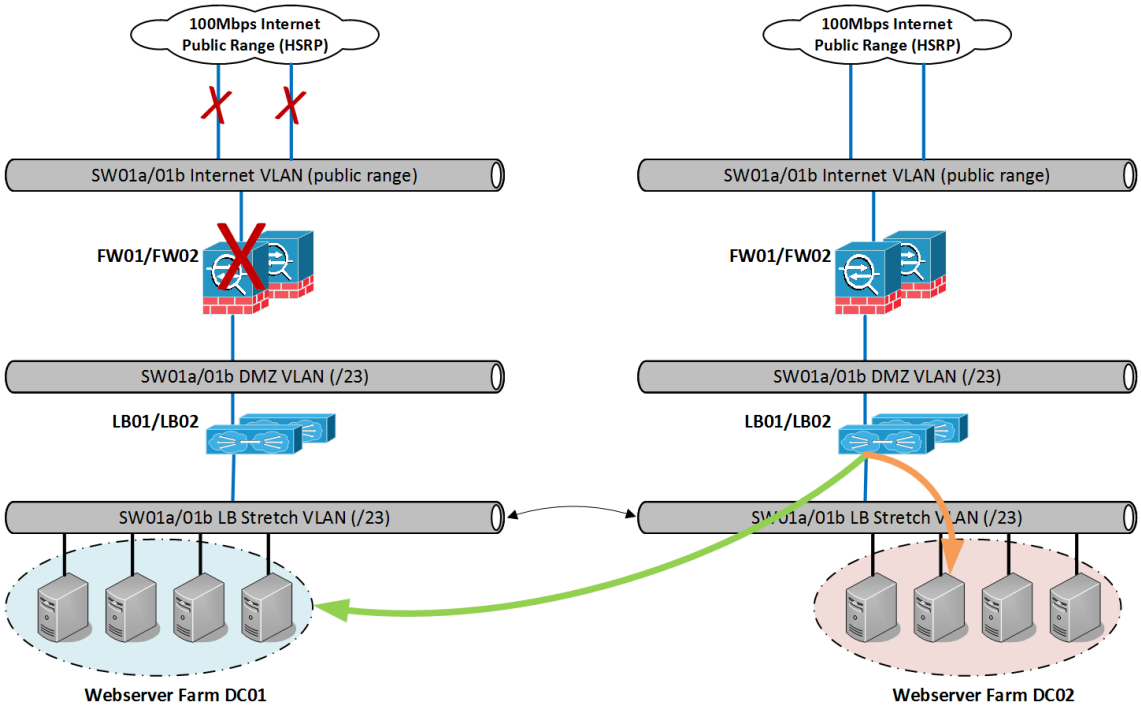

Fig 2.3. Ejemplo de uso de VLANs stretch entre los dos DCs.

A nivel de direccionamiento se ha decidido subdividir la red en dos, asignando la parte inferior al DC primario y la superior al DC secundario. Las máquinas en el DC1 apuntarán a un GW en la primera subred y las máquinas en el DC2 apuntarán a un GW en la segunda subred. Todas ellas usaran la máscara de la superred de manera que la comunicación dentro del a misma VLAN se hará a

en Capa 2. Por ejemplo, si la VLAN stretch es una /23, todas las máquinas del DC1 incluyendo la puerta de enlace estarán en la primera /24 y las del DC2 en la segunda /24.

Esta configuración facilita la comunicación dentro de la misma subred evitando tener que subir el tráfico hasta los firewalls (norte-sur) y encaminarlo a través de una VLAN de tránsito entre los DCs. Con esta arquitectura el tráfico intra-VLAN en los dos DCs (en verde) y el inter-VLAN en cada DC (en azul) funcionará sin problema.

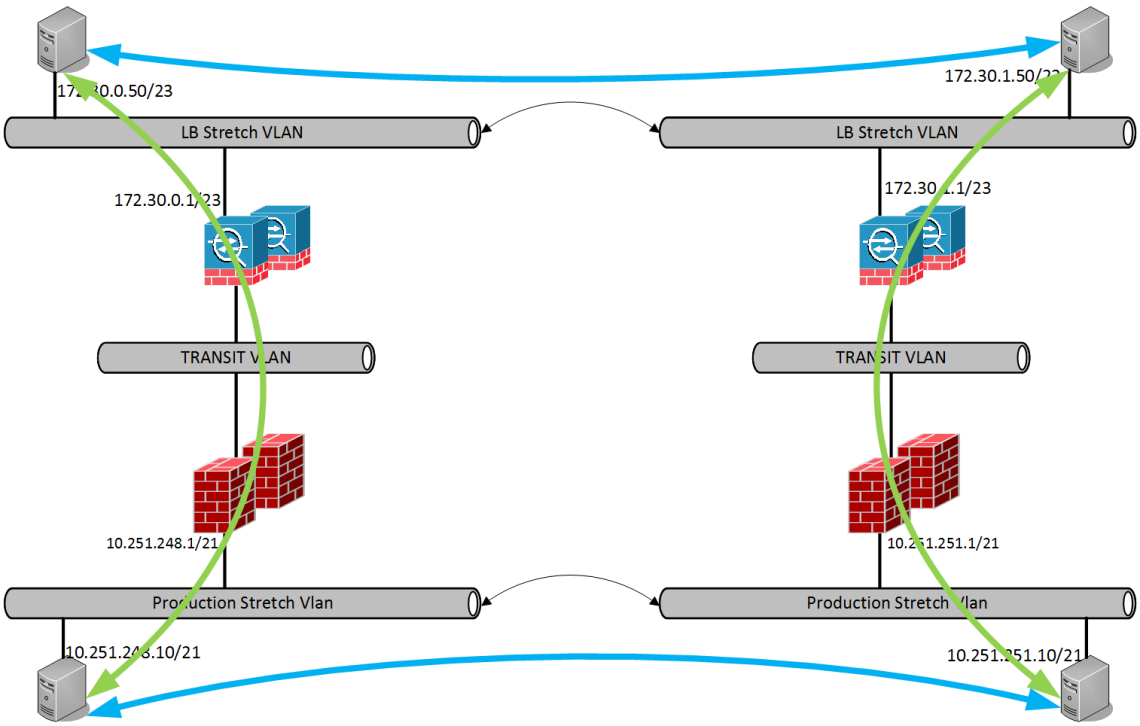

Fig 2.4. VLANs stretch: uso habitual.

2.2.3.1. Problemas de enrutamiento asimétrico

Dada la arquitectura, no se espera que sea habitual el tráfico inter-VLAN inter-DC, pero, de todas formas, será necesario definir reglas para evitar problemas de enrutamiento asimétrico. Para que funcione sin problemas de enrutamiento asimétrico cuando un host en el DC1 quiera conectar con un host en el DC2 en otra VLAN como convención se definirá una ruta estática en la máquina que se encuentre detrás de los firewalls internos (producción) apuntando al firewall en el otro DC.

En el ejemplo de la Fig. 2.5 la máquina 10.251.248.10 (VLAN Prod DC1) quiere contactar con la máquina 172.30.1.50 (LB DC2). Para ello envía el tráfico a su Gateway (10.251.248.1) y este su vez al DC1-FW01 que lo entrega al destino. La máquina de destino, como está en el DC2, tiene como Gateway el DC2-FW01, por lo que los paquetes de vuelta se enviarán hacia allí. El firewall, al recibir un paquete dentro de una conexión que no tiene monitorizada, la descarta.

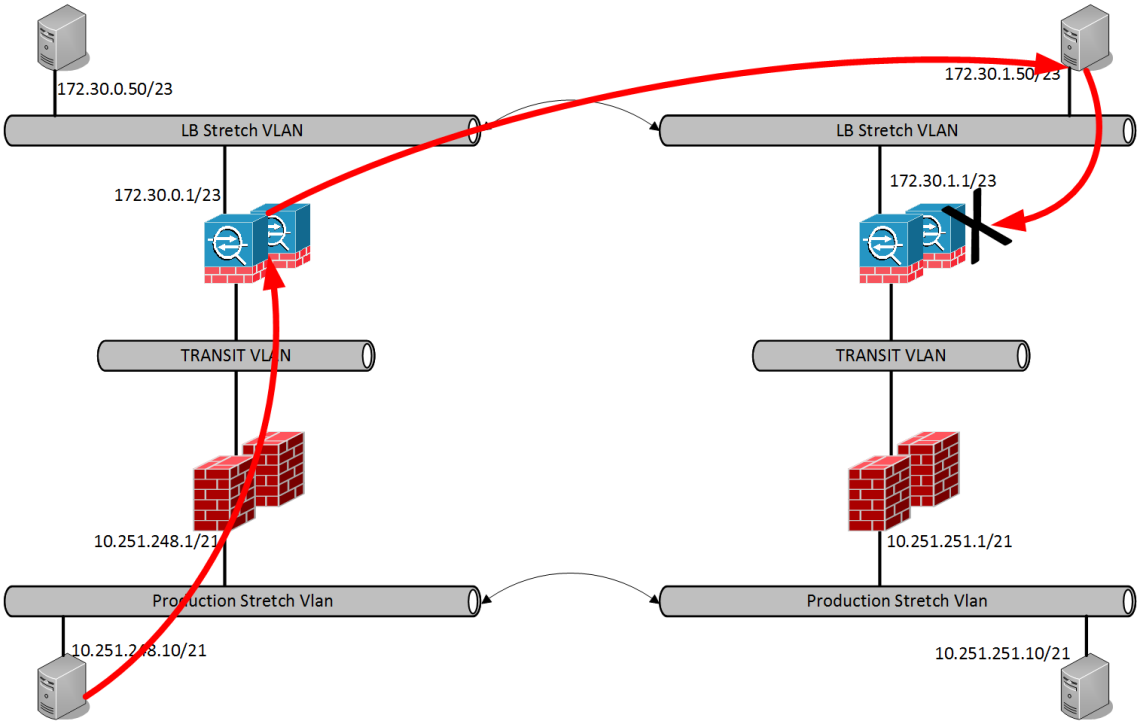

Fig. 2.5. VLANs stretch. Problemas de enrutamiento asimétrico.

2.2.3.2. Enrutamiento asimétrico. Resolución

Para solucionar el problema se añadirá (se espera que esto sea puntual) una ruta estática en la máquina que tiene como Gateway el firewall protegido apuntando al firewall protegido en el DC donde está la máquina de destino. En este caso la máquina 10.254.248.10 tendrá una ruta hacia la subred 172.30.1.0/24 (subred LB en el DC2) apuntando GW en el DC2 para la VLAN de PROD (10.251.251.1). Como podemos ver en la Fig. 2.6, tanto el tráfico de ida como el de vuelta pasarán por los mismos equipos de L3 (firewalls) evitando el enrutamiento asimétrico y permitiendo que los paquetes lleguen en ambos sentidos (y no sean descartados por los firewalls al no pertenecer a una conexión conocida).

Esta casuística, y la necesidad de comunicar inter-VLAN inter-DC, de dará solo de manera puntual para la comunicación entre la VLAN de Producción y la VLAN LB con lo que sólo será necesario añadir una única ruta en las máquinas en la VLAN de Producción que lo necesiten para llegar a la subred LB en el otro DC. Concretamente:

- 172.30.1.0/24 vía 10.251.251.1 para Producción en DC1 que necesiten comunicar con LB en DC2.
- 172.30.0.0/24 vía 10.251.248.1 para Producción en DC2 que necesiten comunicar con LB en DC1.

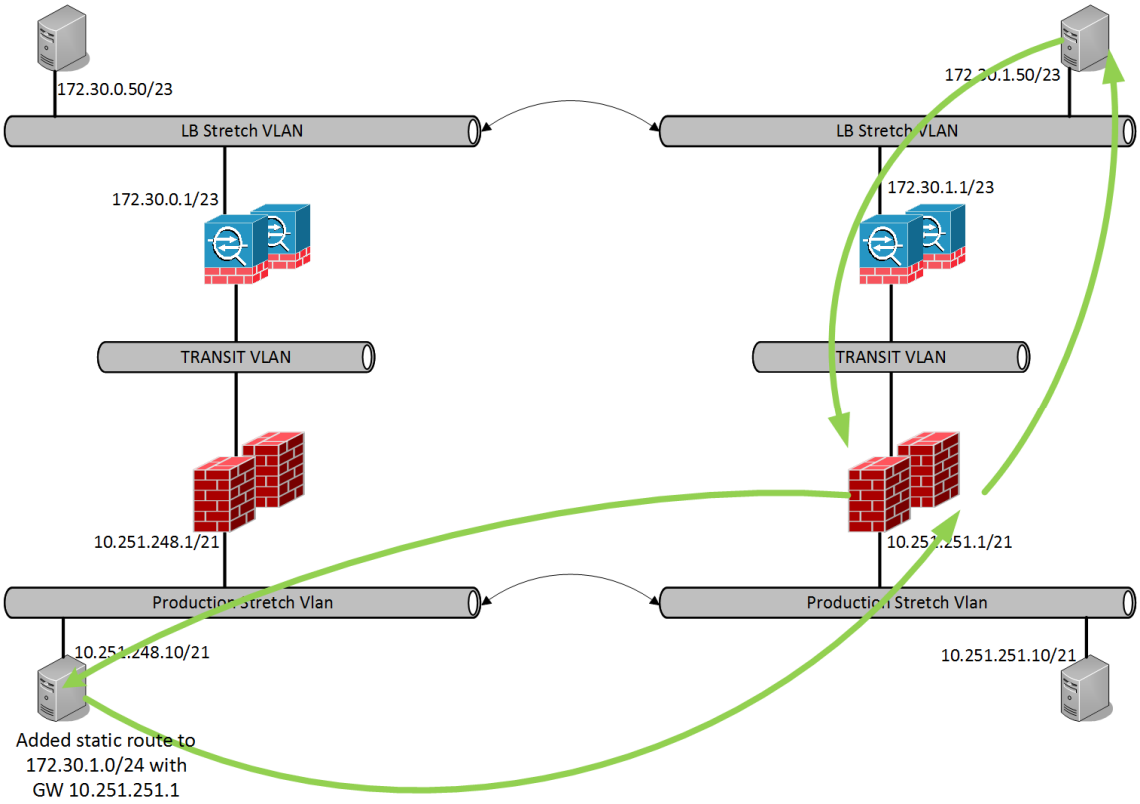

Fig 2.6. VLANs stretch: Solución a los problemas de enrutamiento asimétrico.

Además, esta comunicación (y añadir esta ruta) será necesaria sólo para algunos equipos concretos, ya que en general la comunicación inter-VLAN se da dentro del mismo DC. Un ejemplo son los servidores de dominio (Microsoft Active Directory): habrá uno en cada DC en la VLAN de Producción y todos los equipos en la VLAN LB necesitaran usar estos servidores básicamente por temas de redundancia usando el del otro DC como secundario.

Por lo que respecta al resto de VLANs (las que no son stretch y por tanto son locales a cada DC) no se espera que sea necesaria la comunicación entre DCs. Por si surge esta necesidad en el futuro o de manera puntual, por ejemplo para transferir datos de un DC al otro, se ha definido una VLAN de tránsito entre los DCs, llamada “Inter-DC VLAN”. La configuración, uso y funcionamiento de esta VLAN se define en la sección (2.2.4)

2.2.4. Inter-DC VLAN

La comunicación entre VLANs locales en cada DC (las que no son stretch) no era un requerimiento del cliente. Por si surge esta necesidad en el futuro, aunque sea para transferir un fichero entre ambos DCs, en estas VLANs se propone la utilización de una VLAN de tránsito propagada entre ambos DCs: Inter-DC VLAN.

A diferencia de las stretch VLANs que lo hacen a Nivel 2, esta VLAN de tránsito permitirá la comunicación entre las distintas redes locales de cada DC a Nivel 3 encaminada en los firewalls. Esta VLAN se definirá en los firewalls protegidos (internos), donde están la mayoría de VLANs locales, y se ampliará a los enlaces entre DCs junto al resto de stretch VLANs. En estos firewalls será necesario definir rutas estáticas para llegar a las redes locales del otro DC usando como Gateway el firewall que está en el otro DC. Además, como en cualquier otra comunicación a través de los firewalls, habrá que definir reglas de acceso tanto en el firewall del DC de origen como en el firewall del DC de destino:

- VLAN origen -> Inter-DC VLAN en el firewall de origen.
- Inter-DC VLAN - > VLAN destino en el firewall de destino.

2.2.5. Uso de agregación de puertos.

Entre los distintos switches y, especialmente entre switches y firewalls donde sea posible, se recomienda la utilización de agregación de puertos (LACP o Portchannels) en los equipos Cisco escogidos. Tradicionalmente la conexión de los firewalls a la red se ha hecho utilizando puertos dedicados para los interfaces en los que se espera más tráfico, y algún interfaz en modo trunk para el resto de redes.

En este caso se recomienda utilizar el mismo número de puertos (de manera que la capacidad máxima será la misma) pero utilizando un LACP que los incluya a todos. En el interfaz virtual resultante se transportarán todas las VLANs. Con esto se consiguen dos cosas: Por un lado, que la capacidad para una VLAN concreta podrá ser superior a la de un puerto de manera puntual y por otro, y lo más importante, permitirá la tolerancia en caso de que un puerto concreto falle. En la topología anterior un fallo en un puerto concreto produciría una interrupción del servicio y la activación del failover directamente.

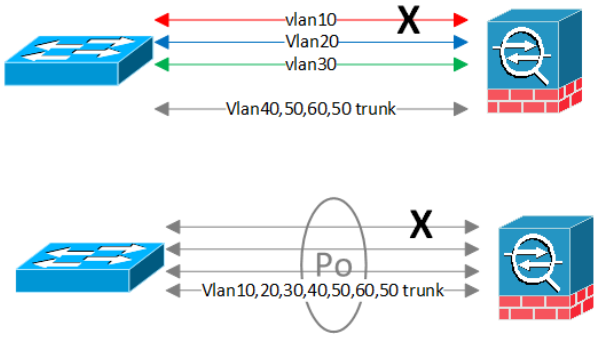

Fig 2.7. Ejemplo de uso de agregación de puertos.

En la Fig. 2.7 se puede ver un ejemplo de esta configuración. En la parte superior tenemos la configuración tradicional y en la inferior la que se propone. Como se puede observar en la figura inferior la VLAN 10 tendrá disponible de manera puntual una capacidad superior a un puerto y, en caso de que un

enlace físico se caiga (X) no se interrumpirá el servicio por esta VLAN (con el consiguiente failover si los firewalls están en alta disponibilidad) sino que el tráfico para esta VLAN seguirá fluyendo por el resto de puertos del LACP. Solo habrá degradación, ya que la capacidad se verá reducida.

2.2.6. Uso de VPCs (Virtual PortChannels).

En la infraestructura serán necesarios caminos redundantes entre los distintos switches para que sea tolerante a fallos tanto en los equipos como en los propios enlaces. Al ser switches independientes la redundancia se realizaría de manera tradicional utilizando STP (Spanning-Tree Protocol). En este tipo de configuración no se pueden utilizar los dos enlaces al mismo tiempo **para la misma VLAN,** ya que se podría formar un bucle. El propio protocolo bloquearía uno de los caminos hasta que cayera el otro.

No es muy óptimo tener enlaces infrautilizados, a la espera de que falle un camino y que STP converja en el otro. Esto es especialmente significativo en los enlaces entre los DCs, ya que tienen un coste asociado bastante alto. Además, también es importante el tiempo de convergencia de STP que, aunque ha mejorado con las distintas versiones del protocolo, sigue sin ser inmediato en algunos casos.

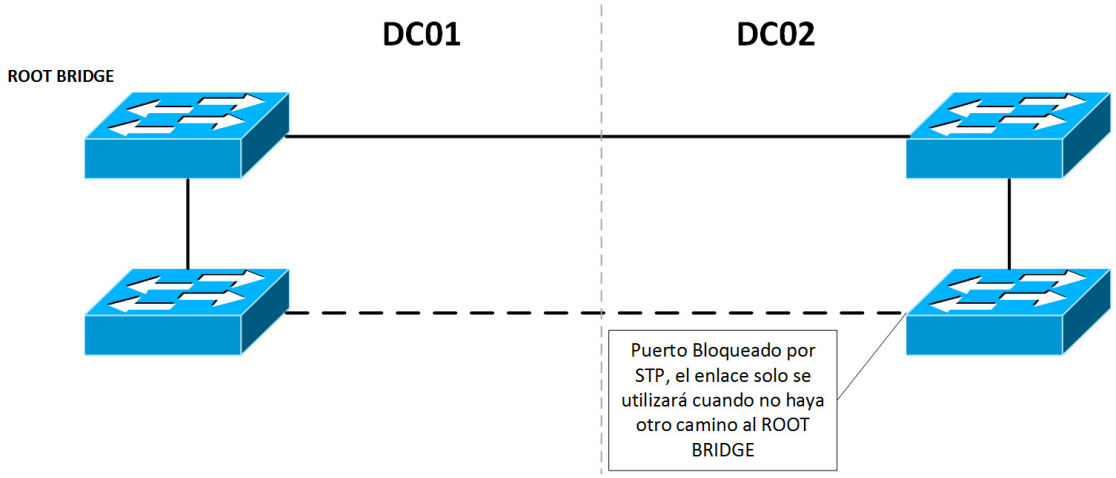

Fig 2.8. Ejemplo de configuración típica entre DCs: un camino activo, el otro bloqueado por STP.

En este caso los switches escogidos para la zona protegida (Cisco Nexus) permiten que enlaces que están físicamente conectados a dos equipos distintos se comporten como un único puerto de capacidad agregada. Esta tecnología se llama vPC (Virtual PortChannel) (ver [3]). Esto permite crear multipathing en capa 2: proporciona redundancia incrementando el ancho de banda, permitiendo múltiples caminos paralelos entre dos equipos y balanceando el tráfico. Otros fabricantes también permiten crear este tipo de configuraciones. En Arista por ejemplo se llama MLAG (ver [4]).

Una vez habilitado en los dos equipos es necesario tener un enlace entre los dos por el que se envían mensajes de hearbeat llamado keepalive-peer-link (normalmente el de gestión) y el peer link. El conjunto de los dos switches el peer-keepalive-link y el peer-link se conoce como vPC domain.

De manera genérica el uso de vPCs proporciona las siguientes ventajas:

- Permite que un único equipo pueda usar un único port-channel contra dos equipos.
- Elimina los puertos (y enlaces) bloqueados por STP.
- Proporciona una topología sin bucles (loop-free topology)
- Permite utilizar todo el ancho de banda agregado de los uplinks
- Proporciona una convergencia rápida en caso de fallo de enlace o equipo.
- Proporciona resiliencia a nivel de enlace.
- En general, ayuda a asegurar la alta disponibilidad.

En la Fig 2.9. podemos ver la propuesta de utilización de vPCs en esta arquitectura:

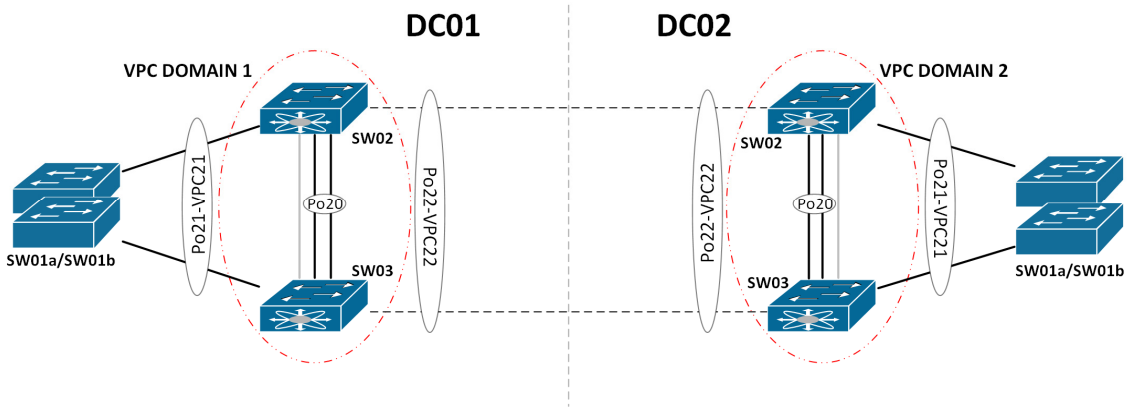

Fig 2.9. Propuesta de utilización de VPCs en la arquitectura

Como se puede ver en la figura Fig 2.9. se propone definir un vPC domain en cada DC entre los dos switches Nexus (SW02/03). Entre ambos equipos habrá por un lado el peer-keep-alive link que corresponde al puerto de gestión (1Gbps, en gris) y el peer-link (Po20) que será un port-channel (2x10 Gbps).

Una vez definido el dominio se crearán dos vPCs: Uno para la interconexión con el switch SW01a/b y otro para interconectar con el otro DC proporcionando así un enlace activo-activo de capacidad agregada. (ver notas)

Por lo que respecta a los switches perimetrales SW01a/SW01b no surge la necesidad de vPC. Los equipos escogidos (Catalysts 3750) están en stack y se comportan como un único switch, con lo que permiten configurar links agregados usando puertos de cualquiera de los miembros.

Nota: Por simplicidad se ha llamado de la misma manera los VPCs en ambos DCs, son independientes y podrían tener IDs distintos.

2.2.7. Interconexión entre DCs – Stretch VLANs, Inter-DC VLAN y uso de vPCs

La interconexión entre los dos DCs se realizará utilizando dos enlaces a 1 Gbps para proveer a la infraestructura de redundancia.

Los enlaces se conectarán respectivamente entre los switches DC1-SW02 y DC2-SW03 y los switches DC2-SW01 y DC2-SW02. Se conectará un de los enlaces a cada par de switches para que el fallo en uno de estos equipos no afecte a toda la conectividad entre los DCs.

Como se ha comentado en la sección 2.2.6, los switches SW02 y SW03 en cada DC estarán en el mismo dominio vPC y se configurará un vPC entre ambos DCs utilizando estos dos enlaces de 1 Gbps proporcionando a la vez redundancia y capacidad agregada de 2 Gbps.

A través de este enlace se transportarán por un lado todas VLANs stretch y por el otro la interDC VLAN en un único trunk.

- Strech VLANs: Proporcionan conectividad a nivel 2 de equipos en la misma VLAN entre los dos DCs.
    - VLAN 40 Customer's Network (office)
    - VLAN 60 LB stretch.
    - VLAN 120 Backup stretch.
    - VLAN 900 Production stretch.
    - VLAN 930 HyperVisor.
- InterDC VLAN: Proporcionan conectividad a nivel 3 entre VLANs en los dos DCs. El enrutamiento se realiza en los firewalls protegidos.
    - VLAN 80 Inter-DC

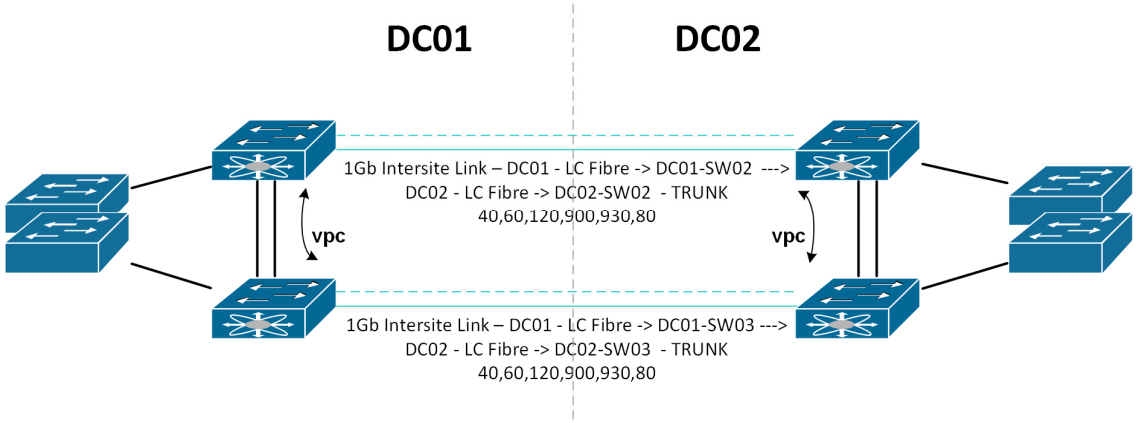

Fig 2.10. Interconexión entre DCs

2.2.8. Conexión las oficinas del cliente y el site de backup

Para la conexión con las oficinas del cliente se utilizará un enlace a 100 Mbps en cada DC conectados a los switches perimetrales. Para asegurar la redundancia en caso de que caiga alguno de estos enlaces se pasará VLAN

del cliente a través del enlace entre los dos DCs y se utilizará STP. Se modificará la prioridad de los puertos para que los dos enlaces estén activos y se utilice el enlace entre DCs en caso de fallo.

Así por ejemplo si falla el enlace contra el DC1 podrán seguir llegando a las máquinas en este DC a través del segundo enlace contra el DC2. Oficina del cliente -> DC2 -> inter-site links -> DC1. (ver figura XXX)

En el caso de los enlaces con el site de backup también habrá dos enlaces a 100 Mbps. pero estos conectados a los switches de la zona protegida. Para ellos se utilizará la misma solución para asegurar la redundancia.

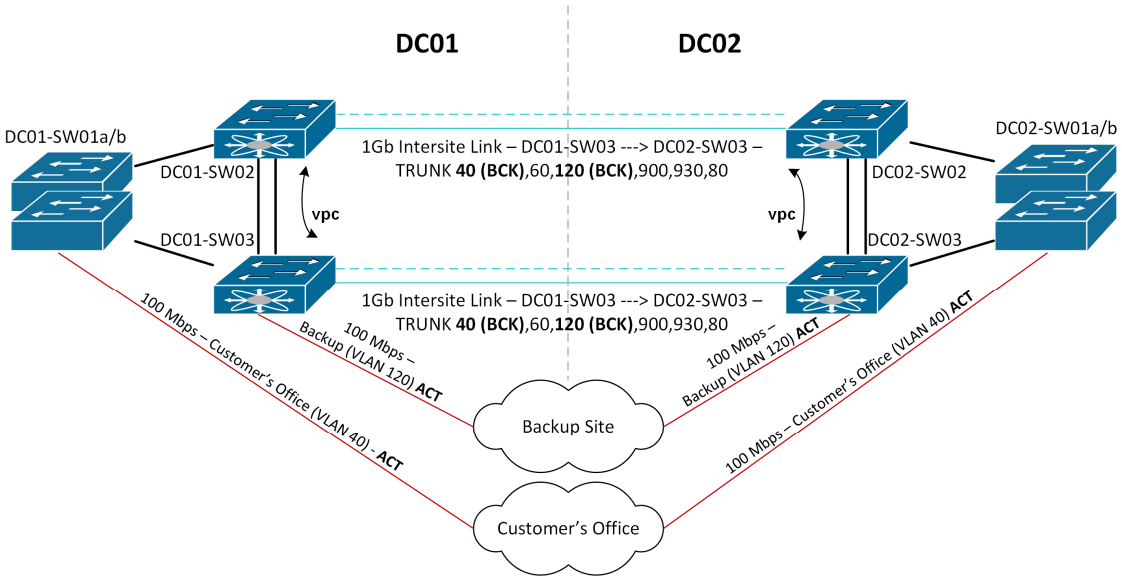

Fig 2.11. Conexión y redundancia para los enlaces con sitios externos.

2.2.9. Acceso a Internet

Se propone la utilización de una salida dual a Internet en cada DC a 100 Mbps. El proveedor proporcionará dos puertos de acceso, conectados a dos CPEs (Customer Premises Equipment) distintos en cada DC. Los dos enlaces se conectarán a los dos switches en la zona perimetral (y por tanto también contra los dos firewalls perimetrales en alta disponibilidad) para asegurar la redundancia.

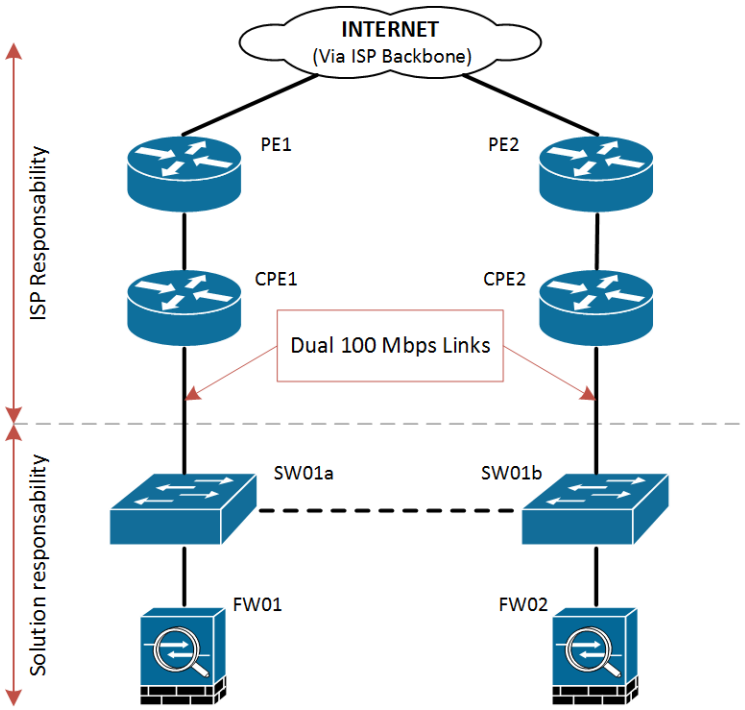

Fig 2.12. Acceso a Internet en cada DC

2.3. Estudio de mercado - Equipos escogidos.

De acuerdo con los requerimientos del cliente y el diseño propuesto, será necesario proveer de los siguientes elementos de red:

- Balanceadores locales.
- Balanceadores globales.
- Firewalls perimetrales.
- Firewalls internos (producción)
- Switches la zona perimetral.
- Switches para el tier de producción (protegido).

2.3.1. Servicio de balanceo

El servicio proporcionará al cliente la capacidad de distribuir la carga de red generada por múltiples sesiones de cliente a lo largo de múltiples servidores. El objetivo del servicio es mejorar la respuesta global de su granja de servidores. Esto se consigue repartiendo la carga de manera que cada uno de los servidores individuales sea capaz de procesar las peticiones que le llegan y asegurando la continuidad del servicio cuando alguno de éstos falla o necesita entrar en mantenimiento.

Estos equipos también descargan los servidores realizando funciones determinadas (normalmente aceleradas por hardware) como pueden ser; la terminación SSL o la gestión de las sesiones de usuarios por ejemplo en

aplicativos como Citrix. Dependiendo del modelo escogido existen integraciones con otras aplicaciones o servicios.

Para esto proyecto, se requieren dos tipos de balanceadores distintos: Balanceadores locales que distribuyan el tráfico entre los servidores de la granja del cliente y balanceadores globales que repartan la carga entre ambos DCs y que sea capaces de mover la carga entre ambos DCs en caso de que uno de ellos falle.

De acuerdo a los requerimientos del cliente (ver 1.1) se ha decidido poner un par de balanceadores locales en cada DC en modo activo/pasivo para balancear las granjas de servidores y asegurar la alta disponibilidad. Así mismo, se ha decidido poner un par de balanceadores globales en modo activo/activo uno en cada DC que se encargarán de balancear los servicios entre los dos DC a nivel de DNS. Según el estado de los servicios en cada momento, estos balanceadores resolverán la IP pública de uno u otro DC.

Se han estudiado los distintos productos en el mercado de distintos fabricantes de acuerdo a los requerimientos del cliente. En este caso se han considerado tanto equipos de la familia Big IP de F5 [6] como de la familia Netscaler de Citrix [7]. Ambos fabricantes son líderes en el mercado (ver [5]).

2.3.1.1. Balanceadores locales.

Los principales requerimientos del cliente para la elección del hardware son: que sea capaz de gestionar un mínimo de 5000 sesiones concurrentes, que sean capaces de realizar la autenticación de servicios como Citrix y que tengan soporte para la terminación SSL.

F5

Entre las soluciones del fabricante F5 encontramos dos modelos que cumplen con estas especificaciones: Big IP 2200s y Big IP 4000s. Ambos son capaces de autenticar servicios como Citrix utilizando el módulo APM, un número mínimo de 5000 sesiones concurrentes a nivel, aceleración y soporte de SSL, 8 puertos Gigabit (con la posibilidad de ampliar los uplinks a 10 Gb con SPFs). Además, ambos equipos permiten la utilización de una fuente de alimentación secundaria facilitando y ayudando a la alta disponibilidad de la plataforma.

Aunque ambos equipos son muy parecidos, el modelo más adecuado para la solución es el 4000s. Las principales razones son que el modelo 4000s permite hasta 10000 sesiones concurrentes (el 2200s llega solo hasta las 5000 que era el mínimo del cliente) y el hecho que por un precio parecido tiene el doble de RAM 16 Gb que permitiría en el futuro la ampliación de los servicios en el equipo con los distintos módulos que ofrece fabricante.

Citrix Netscaler

En lo que respecta a este fabricante, se han comparado los distintos modelos físicos (Netscaler MPX). En este caso los modelos comparados son el MPX-8005 y el MPX-8015. Ambos cumplen los principales requerimientos del cliente:

aceleración y soporte SSL, puertos Gigabit (6 puertos más dos puertos 10 Gb opcionales para los uplinks) y también se permite el uso de una segunda fuente de alimentación. En lo que respecta al soporte de aplicaciones Citrix, lo hace manera nativa, aunque no así el soporte de otras aplicaciones.

Al igual que en el caso de F5 el modelo inferior ofrece un máximo de 5000 sesiones concurrentes, que es el mínimo exigido, por lo que el modelo MPX-8015 sería más adecuado, aunque significativamente más caro que el MPX-8005.

2.3.1.2. Modelo escogido: F5 big IP 4000s.

Después de realizar el estudio de mercado, se ha escogido el modelo Big IP 4000s de F5 ya que es el que más se ajusta a los requerimientos del cliente sin disparar el precio. El modelo inferior de Netscaler era justo respecto a los requerimientos de conexiones concurrentes y el siguiente llegaba a las 15000 (3 veces más) pero a un precio mucho más elevado.

En lo que respecta al soporte de Citrix claramente será mejor (o al menos más ágil) con Netscaler (ya que pertenecen a la misma compañía) pero los equipos F5 lo soportan y además tiene soporte para aplicaciones de otros fabricantes importantes como Microsoft.

Fig 2.13. F5 Big IP 4000 Series

2.3.1.3. Balanceadores Globales

El principal requerimiento del cliente es que los balanceadores sean capaces de balancear utilizando las peticiones DNS de los usuarios. Se prevé que utilicen este servicio 30000 usuarios de manera concurrente y, dado que el protocolo DNS es ligero, el *throughput* de estos equipos no sea de más de 25 Mbps.

2.3.1.4. Equipos escogidos:

Teniendo en cuenta la elección de los equipos BIG IP de F5 para los balanceadores locales, se ha decidido escoger también la solución de este fabricante para los balanceadores globales. La solución está basada en los mismos equipos BIG IP añadiendo un módulo llamado GTM (Global Traffic Manager) [15]. Se ha escogido este equipo principalmente porque de manera

propietaria, los balanceadores globales pueden consultar el estado de los balanceadores locales más allá de comprobar si el servicio está operativo. Utilizando un protocolo propietario pueden saber por ejemplo qué carga tienen o cuantos servidores hay disponibles en cada momento. Esto permite un balanceo de carga global más óptimo.

El módulo GTM de F5 permite realizar el balanceo global con las siguientes características:

- Balanceo global basado en Round Robin, Conexiones, disponibilidad….
- Balanceo basado en métricas de red desde el cliente hasta el DC (por ejemplo, el número de saltos desde el cliente hasta el DC).
- Balanceo basado en información geográfica: La solución incluye una base de datos que identifica la localización en el continente, país y estado/provincia. Esto permite balancear a los usuarios al DC más cercano para asegurar un mejor rendimiento.
- Balanceo personalizado: Se pueden definir otros criterios para balancear hacia un DC u otro.
- Persistencia: Para asegurar que las conexiones hacia las aplicaciones persisten, el equipo mantiene una tabla de persistencia. Así, si un cliente está conectado a un DC, la próxima vez lo enviará al mismo evitando así problemas con las sesiones a nivel de aplicación o SSL.
- Monitorización: De los servicios en cada DC para saber si responden. Además, si los balanceadores locales son Big IP, permite mediante protocolos propietarios, tener información detallada del estado y uso de cada plataforma permitiendo un balanceo más óptimo.

En este caso dado que los requerimientos no son muy altos en cuanto a rendimiento se ha decidido escoger el equipo más pequeño de la familia Big IP, el F5 Big IP 2000s. Aunque sus características superan las necesidades del cliente. Sus principales características se pueden en el anexo 3.

2.3.2. Servicios de seguridad: Firewalls

Los servicios de seguridad comprenden la seguridad tanto a Nivel 2 como a Nivel 3 de red. A nivel dos esta función se lleva a cabo en los switches (ver 2.3.3). A Nivel 3 esto se hace utilizando firewalls.

Las funciones principales de los firewalls son:

- Proteger redes y aplicaciones críticas de ser comprometidas.
- Trabajar para ayudar a prevenir la intrusión por parte de hackers, virus, gusanos y otras amenazas.
- Asegurar la separación de responsabilidades mediante la aplicación de protocolos de gestión.

Los firewalls e IDSs proporcionan protección a la red interna frente a amenazas tanto externas como internas (las más importantes, que muchas veces no se

tienen en cuenta). Esto lo hacen inspeccionando el tráfico entre todas las máquinas que se comuniquen a través de ellos bloqueando o permitiendo el tráfico de acuerdo a distintos factores: IP, protocolo de transporte, puerto…

De acuerdo a los requerimientos del cliente (1.1) se propone la utilización de dos niveles de firewalls, uno para la zona perimetral y otro para la zona protegida. Con el fin de asegurar el estándar EAL4 (Evaluation Assurance Level 4) [8] se utilizarán dos fabricantes distintos. Con esto se consigue por ejemplo minimizar la posible incidencia de un bug descubierto en un equipo en concreto.

El hecho de utilizar dos niveles de firewalls proporciona también protección adicional para los equipos en la zona protegida frente a ataques externos. Cualquier comunicación desde el exterior tendrá que pasar por los dos niveles de firewalls.

Para asegurar la alta disponibilidad se propone utilizar dos de firewalls en modo activo/pasivo en la zona perimetral y otros dos firewalls en el mismo modo en la zona protegida. La misma configuración se utilizará en los dos DCs.

Aunque los requerimientos a nivel de switch eran tener puertos a 1 Gbps en la zona perimetral y a 10 Gbps en la zona protegida, se espera que el tráfico de mayor volumen (especialmente tráfico de replicación entre servidores) se dé dentro de las mismas VLANS, a Nivel 2. Por los firewalls sólo pasará el tráfico entre las distintas VLANs públicas y privadas del cliente. Se estima que el tráfico en el firewall perimetral no superará los 600 Mbps y en el firewall protegido los 2 Gbps. Para cumplir con estos requerimientos se han estudiado equipos de varios fabricantes, incluyendo Cisco, Checkpoint, Fortinet y Juniper.

2.3.2.1. Firewall perimetral

Para la zona perimetral se ha escogido la plataforma Cisco ASA [9]. Los equipos de esta familia proporcionan funcionalidades de firewall básicas, pudiendo definir reglas utilizando la IP de origen, IP de destino y puerto. Además, proporcionan lo que se conoce como *stateful inspection* de más de 30 aplicaciones distintas permitiendo la inspección a nivel 4-7 de estos protocolos. Estos equipos pueden ser configurados en alta disponibilidad para asegurar el funcionamiento en caso de fallo de un equipo.

Además, la última generación de los firewalls Cisco ASA están preparados para trabajar con la tecnología NGIPS (Next-Generation IPS) llamada Firepower [10] [11] sólo añadiendo licencias adicionales. Con esto, se cumpliría el requerimiento del cliente de proveer de una solución de IDS.

2.3.2.2. Equipo escogido: Cisco ASA 5515-X

En este caso se ha escogido el modelo Cisco ASA 5515-X ya que es el que más se ajusta a los requerimientos en cuanto a *throughput* y conexiones

concurrentes. Las principales características de este equipo están detalladas en el Anexo B.

Fig 2.14. Cisco ASA 5515-X firewall

Se utilizarán un par de estos equipos en cada localización. Cada uno de ellos tendrá conectividad directa con Internet, la red del cliente en sus oficinas, las VLANs de DMZ y LB y el resto de redes detrás del firewall protegido a través de la VLAN TRANSIT. Además, desde ellos se configurarán VPNs LAN-to-LAN o de acceso remoto con los distintos proveedores del cliente.

2.3.2.3. Firewall protegido

Para esta capa secundaria de firewalls se propone la utilización de firewalls del fabricante Checkpoint [12], uno de los fabricantes más reconocidos de equipos de seguridad [13] En este caso también se propone la utilización de un par de firewalls en cada DC en HA. En este caso la alta disponibilidad se da utilizando el protocolo estándar VRRP (o el propietario de checkpoint Cluster XL [14]).

2.3.2.4. Equipo escogido: Checkpoint 4600

De acuerdo a los requerimientos se ha escogido el equipo Checkpoint 4600. Sus principales características se encuentran en el anexo C

Fig 2.15. Checkpoint 4600 Series firewall

2.3.3. Servicios de red de área local (LAN)

Los servicios de red de área local proporcionan la conectividad Ethernet/IP utilizando switches. En estos equipos también se proporcionan características de seguridad, VLANs y QoS avanzado si fuera necesario mientras se mantiene la simplicidad del switching tradicional.

Los requerimientos básicos del cliente y que se han tenido en cuenta a la hora de escoger estos equipos son:

- Conectividad al menos a 1 Gbps en el tier perimetral.
- Conectividad a 10 Gbps para el tier protegido (producción). (16 puertos x 3 *blade chassis*)
- Alta disponibilidad a nivel de equipo. (N+1) en cada DC.

Además, se ha tenido en cuenta la capacidad de los equipos en cuanto a redundancia (fuentes de alimentación, redundancia a nivel de equipo) y, ya que los equipos de red con frecuencia se reutilizan en las sucesivas renovaciones de las plataformas, las capacidades en cuanto a DC del futuro (SDN, transporte/encapsulación de VXLANs…).

De acuerdo a estos requerimientos se han considerado equipos de distintos fabricantes centrándonos específicamente en los fabricantes Cisco, que es líder del mercado con las familias de switches Catalyst y Nexus, así como de Arista, que actualmente está ofreciendo productos que ofrecen una buena calidad -precio en entornos de DC.

2.3.3.1. Zona perimetral

Cisco

Por lo que respecta al fabricante Cisco entre las múltiples familias de switches se han considerado los modelos 3560-X y 3750-X [16]. Ambos equipos cumplen el principal requerimiento respecto a la conectividad, tienen 48 puertos a 1 Gbps. También ofrecen hasta 4 uplinks a 10 Gbps que se utilizarían para interconectarlos con los switches en la zona protegida y redundancia a nivel de fuente. Además, colocando 2 en cada DC se consigue también redundancia N+1 en caso de la caída de uno de los equipos.

La principal diferencia entre estas dos opciones es que los equipos de la serie 3560-X son switches independientes mientras que la serie 3750-X permite conectar los equipos en Stack [17]. Interconectados con el conector stack-wise se comportarían como un único equipo de capacidad agregada: La gestión es única y permiten agregar puertos de ambos equipos como si fueran un único enlace (etherchannels [18]) consiguiendo así una mayor redundancia. Otra ventaja es que no se desaprovechan puertos de red para la interconexión de los equipos.

Arista

Por lo que respecta a Arista, el fabricante está más centrado en equipos de backbone y datacenter de alta capacidad. En este caso el único equipo que se ajusta a los requerimientos para esta zona es el 7010T-48 [18]. Estos equipos son independientes (Standalone) y ofrecen 48 puertos a 1 Gbps y la posibilidad también de ampliar con hasta 4 uplinks de 10 Gbps SPF+.

Aunque no se pueden configurar en stack, el fabricante tiene una tecnología propietaria que permite agregar puertos (LACP) de múltiples equipos. Esta tecnología se llama MLAG [4].

2.3.3.2. Zona protegida

Cisco
Por un lado, dentro de la familia Catalyst se ha estudiado el modelo Serie 4500 [19] ya que es el único que ofrece la posibilidad de utilizar puertos a 10 Gbps. Usando estos equipos es necesario reservar al menos un módulo para la supervisora. Los módulos de puertos 10 Gbps tienen cada uno 12 puertos por lo que mínimo, sería necesario utilizar el chasis 4506E que dispone de 6 módulos (supervisora + 5 para puertos).

Otra opción que se ha considerado es utilizar los nuevos equipos Nexus especialmente pensados para entornos de datacenter. En concreto el modelo 5548UP [20] cumpliría con los requerimientos del cliente en cuanto a densidad y capacidad de los puertos; ofrece 32 + 12 puertos SPF+ que permiten conectividad a 10 Gbps, así como redundancia a nivel de fuente. Además, utilizando la tecnología vPC [3] permite agregar puertos pertenecientes a distintos equipos aprovechando la capacidad de todos los uplinks y evitando utilizar STP entre estos equipos.

Otra ventaja importante en comparación con los equipos Catalyst es que estos equipos son mucho más compactos (1 U) con el consiguiente ahorro económico en espacio en el DC. Además, y ya de cara a futuras actualizaciones, permiten transportar también tráfico de storage (FiberChannel) así como desencapsular paquetes VXLAN en futuras redes SDN.

Arista
Por lo que respecta al fabricante Arista hay varios modelos que cumplen con los requerimientos en cuanto a capacidad del cliente, en concreto cabe destacar la serie 7050X [21]. Dentro de esta, el fabricante ofrece distintos modelos dependiendo la cantidad y tipos de puerto. En concreto el modelo 7050SX64 sería el más adecuado para esta implementación. El equipo es muy parecido al Cisco Nexus del apartado anterior: compacto en una U, permite conectar hasta 64 puertos 10 Gbps utilizando conectores SPF+ y tiene redundancia a nivel de fuente.

Además, también permite la agregación de puertos de distintos equipos utilizando la tecnología propietaria MLAG y también está preparado para futuras ampliaciones permitiendo por ejemplo transportar VXLANS. Estos equipos son puramente de red y no permiten transportar tráfico de storage nativo (FiberChannel).

2.3.3.3. Modelos escogidos: Cisco Catalyst 3750-X y Cisco Nexus 5548-UP

A la hora de escoger los switches entre todos los presentados en la sección anterior se ha decidido que todos los equipos fueran del mismo fabricante para evitar problemas derivados de la utilización de protocolos propietarios, así como para facilitar la gestión. Aunque todos los equipos expuestos podrían utilizarse se ha decido escoger Cisco ya que es el fabricante que ofrece equipos para ambas zonas más ajustados a los requerimientos.

Para la zona perimetral se ha optado por escoger equipos 2 x Cisco Catalyst 3750 con 48 puertos a 1Gbps y 1 puerto a 10 Gbps, en configuración N+1. Estos equipos se utilizarán para la conectividad de todos los equipos en este Tier: Firewalls externos, Balanceadores, GTMs así como distintos servicios de conectividad para el cliente: enlaces hacia las oficinas de los clientes o los accesos a internet. También se utilizarán para la conexión de los interfaces de gestión de toda la infraestructura.

Fig 2.16. Cisco Catalyst 3750-X

La zona protegida o de producción se usará especialmente para la integración de los servidores del cliente. En este caso se trata de *blade chassis* de alta capacidad por lo que el principal requerimiento para estos equipos será la posibilidad de trabajar con velocidades de puerto a 10 Gbps, así como la posibilidad de configurar *trunks* en los puertos hacia los servidores.

Para el tier de producción se ha decidido utilizar switches Nexus 5548P configurados también en N+1 en cada datacenter. Cada switch tendrá 32 puertos a 10 Gbps y un módulo de expansión con puertos a 1 Gbps. Cada uno de los puertos vienen vacíos por lo que se puede escoger un SFP+ u otro (1 o 10 Gbps) dependiendo de las necesidades.

Fig 2.17. Cisco Nexus 5548-UP

2.4. Propuesta de arquitectura física

Teniendo en cuenta el diseño inicial de la solución, las consideraciones de diseño realizadas y a los equipos escogidos, en la Fig. 2.18, se presenta la propuesta de arquitectura física para el cliente a bajo nivel. En la figura se puede ver de manera detallada todos los enlaces, equipos y tecnologías utilizadas en esta arquitectura para asegurar la alta disponibilidad y la redundancia cumpliendo así con todos los requisitos que se habían planteado (ver 1.1). Las consideraciones más importantes de este diseño se explican a continuación.

Como se puede ver en la figura, en cada DC tenemos dos zonas diferenciadas: Por un lado, la zona perimetral y por el otro lado la zona protegida.

El elemento principal de conectividad en la zona perimetral son los switches Catalyst SW01A/B que interconectan los firewalls, LBs y servicios de red (salida de internet, conexión con las oficinas del cliente) en esta zona con los switches en la zona protegida.

Para conectar los firewalls perimetrales (Cisco ASA) se ha utilizado agregación de puertos para pasar las distintas VLANs del cliente y un puerto adicional que se utilizará para asegurar el intercambio de información entre los equipos en caso de failover.

Por lo que respecta a los balanceadores (F5 Big IP), se ha utilizado un puerto para cada una de las VLANs del cliente y también uno adicional para el failover.

La salida a Internet se ha conectado a cada uno de los switches para garantizar la alta disponibilidad.

Utilizando enlaces a 10Gbps en cada uno de los switches y agregación de puertos estos se conectan a los switches de la zona protegida.

El elemento principal en la zona protegida son los switches Nexus (SW02/03). Utilizando puertos a 10 Gbps estos interconectan con los switches perimetrales y con los blade servers del cliente donde estarán alojados los servidores utilizados en esta solución. Además, utilizando puertos a 1 Gbps interconectarán los firewalls protegidos (Checkpoint 4600), los enlaces entre los dos DCs y los enlaces hacia el site de backup.

Para la comunicación con los firewalls Checkpoint se utilizarán distintos puertos por los que se pasarán utilizando múltiples trunks todas las VLANs protegidas por los equipos, así como la VLAN que utilizan para replicación en caso de failover.

Para interconectar los blade chassis del cliente se han utilizado múltiples puertos a 10 Gb (8 por chasis y por switch) consiguiendo una topología totalmente mallada. Estos puertos se agregarán en distintos LAG (link aggregation port) o Portchannels por los que se pasara un trunk con todas las VLANs presentadas hacia los servidores.

Finalmente, por lo que respecta a los enlaces entre los DCs, se conectará cada uno a uno de los switches para conseguir redundancia y, utilizando VPCs se agregaran consiguiendo tener así un enlace virtual de capacidad agregada (2 Gbps). Por ellos se pasarán todas las VLANs extendidas entre los dos DCs (stretch), la interDC VLAN para tener conectividad en capa 3 entre las dos plataformas y las VLANs de backup y de las oficinas del cliente, para tener redundancia en caso de que alguno de los enlaces hacia estas localizaciones falle.

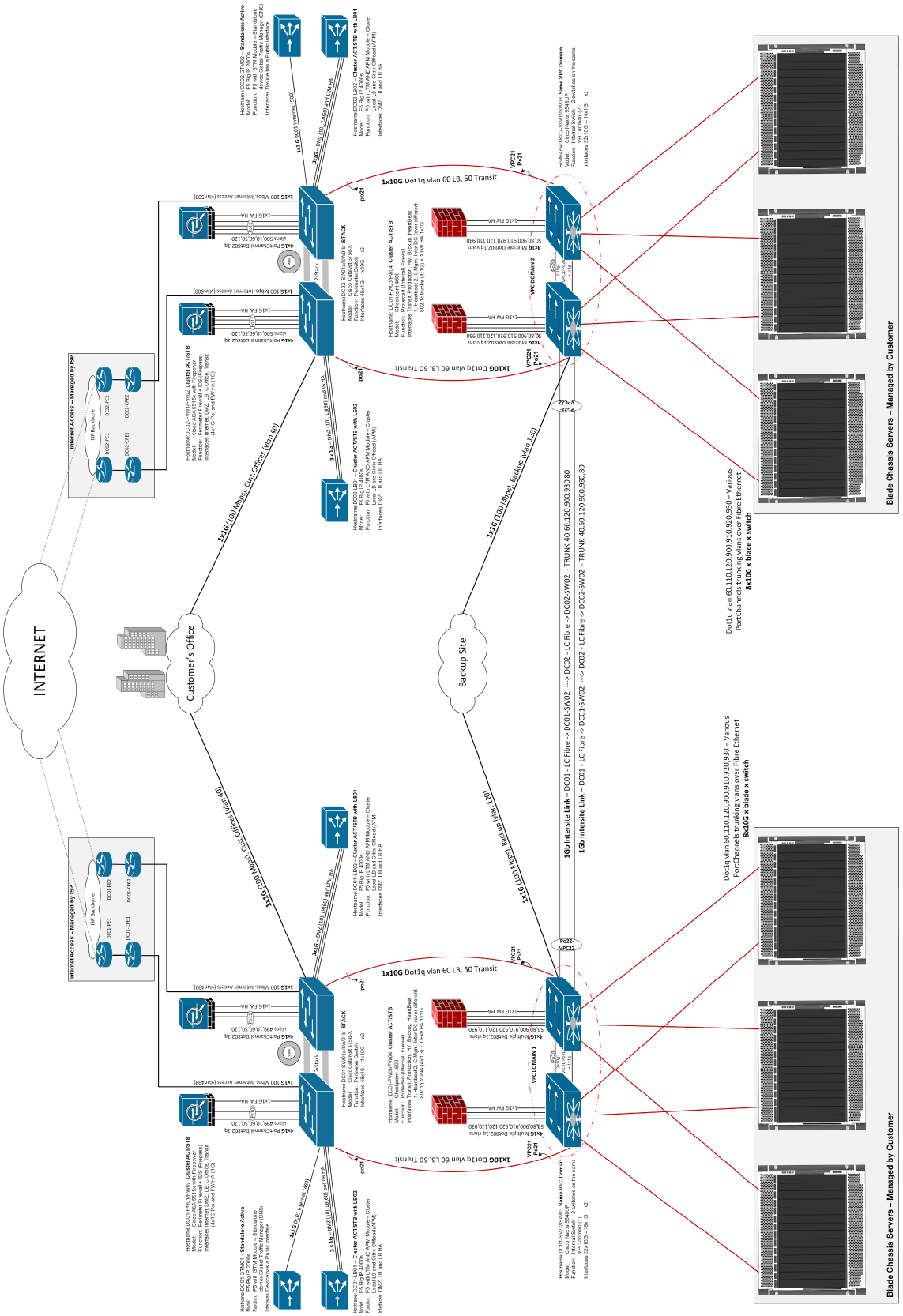

Fig 2.18. Propuesta de arquitectura física

CAPÍTULO 3. PROPUESTA DE DISEÑO BASADA EN VIRTUALIZACIÓN DE RED

3.1. Introducción

De manera alternativa al diseño propuesto, consistente básicamente en una arquitectura de red física, se propone la implementación de la solución para el cliente utilizando una arquitectura virtual basada en SDDC (Software Defined DataCenters) y SDN (Software Defined Networks), tecnologías que actualmente se están empezando a utilizar para substituir el uso de arquitecturas tradicionales en DCs. El objetivo es comprobar si es posible satisfacer los requerimientos de un cliente real y ver qué ventajas e inconvenientes presenta el uso de estas tecnologías, a la hora de hacer una implementación de este tipo.

Para el diseño de esta arquitectura se ha decidido utilizar la solución NSX [22] de VMware para la virtualización de la red, junto con la virtualización de servidores que desde hace años propone este fabricante. En el anexo D se explican los componentes y funcionamiento de esta arquitectura y en el anexo E los requerimientos de la red física subyacente, así como cómo hacer el despliegue de NSX sobre ésta.

El principal requerimiento para conseguir una total integración es que el cliente utilice VMware como Hipervisor para sus máquinas virtuales.

Después de estudiar la arquitectura NSX se ha podido comprobar que está especialmente indicada para grandes despliegues en los que se necesita gran escalabilidad como los de un proveedor de servicios gestionados o de Cloud. Especialmente por los costes asociados al despliegue inicial de la arquitectura de red subyacente y de NSX, hacer un despliegue completo para una solución como la del cliente no sería realizable, ya que estos no se podrían compensar con las ventajas que ofrece.

En este proyecto se propone utilizar la infraestructura de NSX ya desplegada por un proveedor de servicios compartida con sus distintos clientes para desplegar la solución del cliente encima pudiendo aprovechar así las ventajas que ofrece la virtualización de red y, a la vez, la reducción de costes que ofrece utilizar una infraestructura compartida.

En este capítulo se explica la propuesta de diseño basada en virtualización de red para la solución del cliente. Para ello en la sección 3.2 se explican las características del despliegue NSX y la red física subyacente utilizada por el proveedor y en las secciones 3.3 y 3.4 se expone la propuesta de arquitectura virtual para el cliente y las consideraciones de diseño que se han tomado para realizarla respectivamente.

3.2. Arquitectura NSX y red física subyacente

Aunque no es objetivo de este proyecto en esta sección se ofrece una descripción a grandes rasgos de la infraestructura NSX del proveedor de servicios en los dos DCs sobre la que se desplegará la solución de red virtualizada para el cliente.

Para la implementación de esta solución se propone utilizar la infraestructura de NSX ya desplegada por un proveedor de servicios compartida para sus distintos clientes. Se asume que el proveedor tendrá dos infraestructuras de NSX independientes, una en cada DC, de acuerdo con las recomendaciones del fabricante (Anexo E). Éstas incluyen tanto el despliegue de NSX, como la red física subyacente y estarán interconectadas para poder ofrecer servicios en alta disponibilidad entre DCs.

Como se puede ver en la siguiente figura en cada uno de los DCs habrá una serie de clústeres de servidores ESXi de VMware donde se desplegará la solución NSX separados en distintos racks por funciones. Estas funciones son:

- Management Racks: En estos racks se hospedarán los servicios de gestión de la infraestructura; vCenter server (para gestionar la virtualización de los servidores), NSX Manager y NSX controllers.

- Compute racks: Es donde se desplegarán las máquinas virtuales de los clientes.

- Hasta ellos solo se extenderá la VLAN utilizada para transportar VXLANs y otras VLANs de gestión.

- Edge racks: En estos racks es donde se realiza la interacción entre las redes físicas y la red virtual montada sobre NSX. En ellos se realizarán las siguientes funciones

- Se dará conectividad entre la red física exterior y la red virtual. En el clúster de servidores ESXis en estos racks se desplegarán los NSX Edge Gateways appliances que harán de Gateway L3 hacia la red física.
- Se dará conectividad contra equipos físicos conectados a VLANs en la red física. En este clúster se desplegarán las instancias que realizan las funciones de L2 bridge entre las VXLANs lógicas y las VLANs físicas.
- Se centralizarán los servicios de red ya sean virtuales o físicos (que se vayan a integrar con la infraestructura NSX. Pueden ser equipos como balanceadores, firewalls, IDS…
- Todos los servidores ESXi en este clúster tendrán visibilidad de las VLANs físicas en el DC que tengan que tener conectividad contra la infraestructura NSX: VLANs conectadas contra los routers que dan acceso a Internet/WAN, VLANs que se tengan que integrar con las VXLANs en la infraestructura NSX, VLANs que conecten con localizaciones remotas usando servicios en Capa 2/3.

Para interconectar todos estos racks se utilizará una topología de red leaf-spine. En cada uno de los racks habrá un par de switches leaf. Estos harán de Gateway para las VLANs locales en cada rack y conectarán en capa 3 con una

serie de switches spine. Entre las capas leaf y spine se utilizará enrutamiento dinámico y Equal-Cost Multipathing (ECMP) para utilizar todos los caminos de manera simultánea sin crear bucles. NSX desplegará la red lógica L2/L3 sobre esta arquitectura de manera transparente.

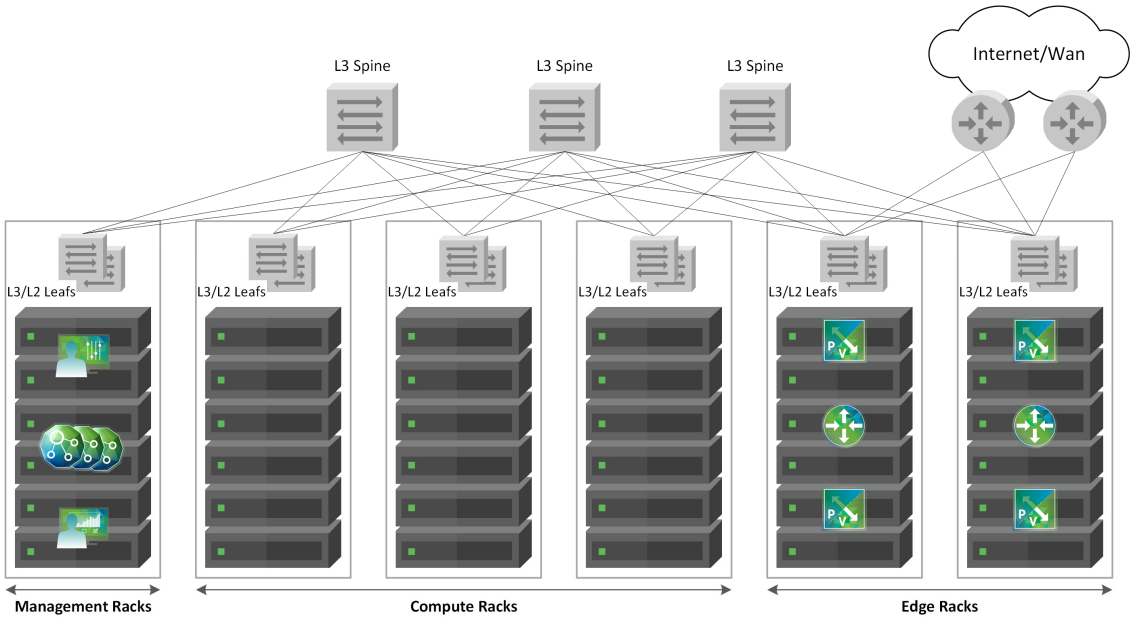

Fig 3.1. Arquitectura NSX y red física en cada DC.

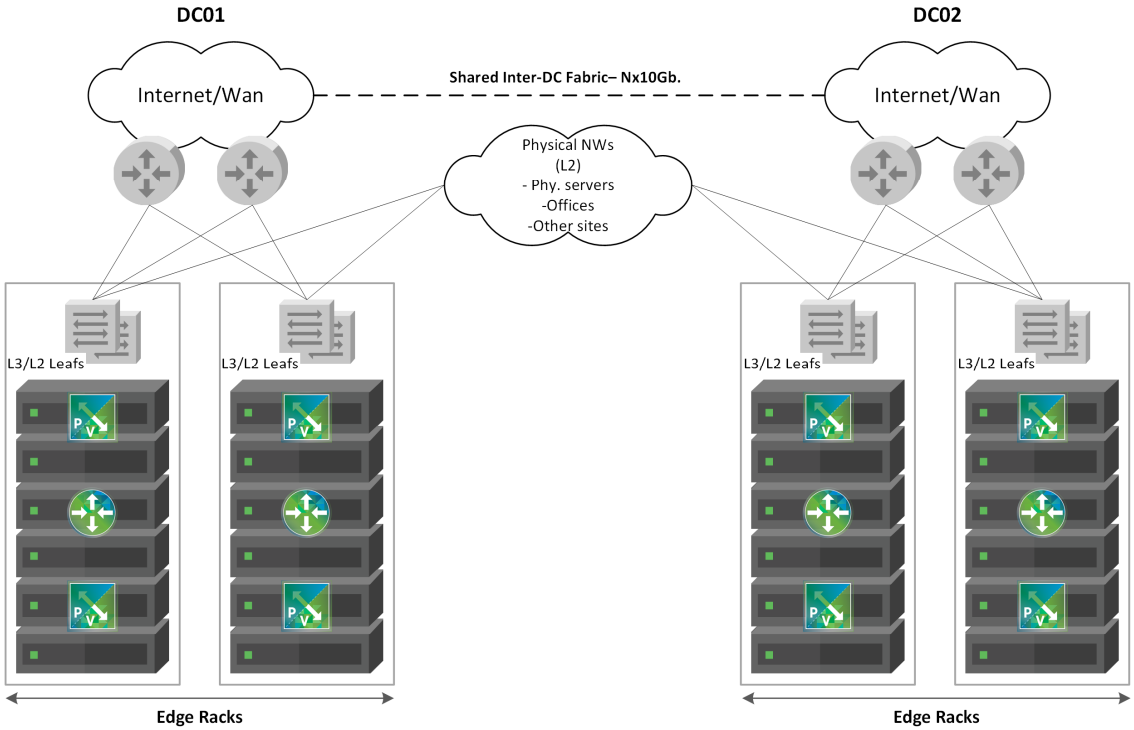

Fig 3.2. Detalle de los Edge racks e interconexión entre DCs

En la Fig 3.2. se puede ver el detalle de los Edge Racks y la interconexión de los dos DCs. Para la interconexión el proveedor tendrá una seria de links agregados que se compartirán con todos los clientes de la plataforma.

En estos racks se conectarán también todas las redes físicas que se tengan que integrar con el dominio virtual. Estas se extenderán a los equipos dentro del rack donde se hace el bridging tanto en capa 2 como en capa 3. En el caso de la infraestructura del cliente aquí se conectarán los enlaces a sus oficinas y el site de backup así como la salida a Internet.

3.3. Propuesta de arquitectura virtual

En esta sección se presenta una propuesta de diseño basada en virtualización de red para la solución del cliente, desplegada sobre una solución de NSX. Esta propuesta se puede ver en detalle en la Fig.3.3.

En la parte superior de la figura se puede ver el diseño de la solución virtual para el cliente. Este diseño se ha basado en la propuesta de diseño físico hecha en el capítulo 2. Lo que se pretende es, a nivel funcional, ofrecer el mismo servicio. Para ello se han substituido los distintos elementos de red físicos por los elementos de la arquitectura NSX que permitan ofrecer funciones equivalentes. Algunos de estos elementos se ofrecerán de manera centralizada utilizando appliances virtuales o de manera distribuida, en el kernel de los servidores físicos que componen la solución.

Al igual que en la arquitectura física, este diseño está basado en dos zonas, una perimetral y una protegida, distribuida en dos DCs. Los dos DCs estarán interconectados para asegurar la replicación entre ambas plataformas y cada uno de ellos lo estará con las oficinas del cliente y el site de backup. Además, cada DC tendrá una salida a Internet dual a 100 Mbps en la red física, que estará conectada con los equipos que hacen de pasarela entre la red física y virtual en la arquitectura. Cada uno de los DCs dispone de una arquitectura NSX independiente con sus propios NSX managers y NSX controllers.

Lo primero que se puede observar es que todas las VLANs del cliente pasan a estar gestionadas por el switching virtual de VMware/NSX (vNetwork Distributed Switch - vDS). Los paquetes de cada una de ellas serán encapsulados utilizando VXLAN para poder ser transportados a través de la red subyacente de manera transparente.

Por lo que respecta a la zona perimetral, en ella están conectadas todas las VXLANs que necesitarán tener acceso directo a la red física de Internet, esto incluye la DMZ del cliente y la red donde los balanceadores publican los servicios. En esta zona también estarán conectados los balanceadores globales para balancear la carga entre los dos DCs. Para ofrecer estos servicios se utilizarán:

- 2 x NSX Edge Gateways en alta disponibilidad que se encargarán de interconectar las redes Internet, DMZ y LB. Además, estarán conectados

a la red de transito que da acceso a la zona protegida. Estos equipos englobarán múltiples funciones: se encargarán del enrutamiento de estas redes, de hacer de firewall entre ellas (incluyendo el NAT y las VPNs hacia los proveedores) y de los servicios de balanceo local entre los servidores de la granja del cliente.

- 1 x Balanceador Global. Como el servicio de balanceador global no estaba disponible en la arquitectura, se utilizarán los equipos F5 utilizados en la arquitectura física en su versión virtual, directamente conectados a la red pública. Se utilizará uno en cada DC en modo activo-activo.

Por lo que respecta a la zona protegida, en ella estarán todas las redes criticas del cliente que no serán accesibles directamente desde Internet. Para esta zona se han utilizado:

- VDR (Virtual Distributed Routers) para el enrutamiento entre las redes conectadas.
- DFW (Distributed Firewalls) para proporcionar seguridad en la comunicación entre las redes de la zona y para ofrecer una segunda capa de seguridad para las conexiones que vengan desde Internet a través de la zona de tránsito.

Como se ha comentado, dentro de la arquitectura del cliente todas las redes estarán virtualizadas. En algunos casos serán necesario conectar redes físicas que no están en el dominio NSX. Para ello se han utilizado:

- 2 x NSX Edges. Habilitados como L2 VPNs para transportar las VXLANs que estarán extendidas entre las dos plataformas a través de los enlaces compartidos del operador entre los dos DCs.
- L2 bridge (distribuido). Esta funcionalidad se ha utilizado para conectar la oficina del cliente y el site de backup que se encuentran en el dominio físico con su VXLAN correspondiente dentro de la arquitectura del cliente. Los enlaces físicos hacia estas dos localizaciones estarán conectados a la infraestructura física del DC y la VLAN correspondiente se presentará a los L2 bridge.

En la parte inferior de la Fig.3.3. se puede ver de manera esquemática la arquitectura física sobre la que se montará la solución virtualizada.

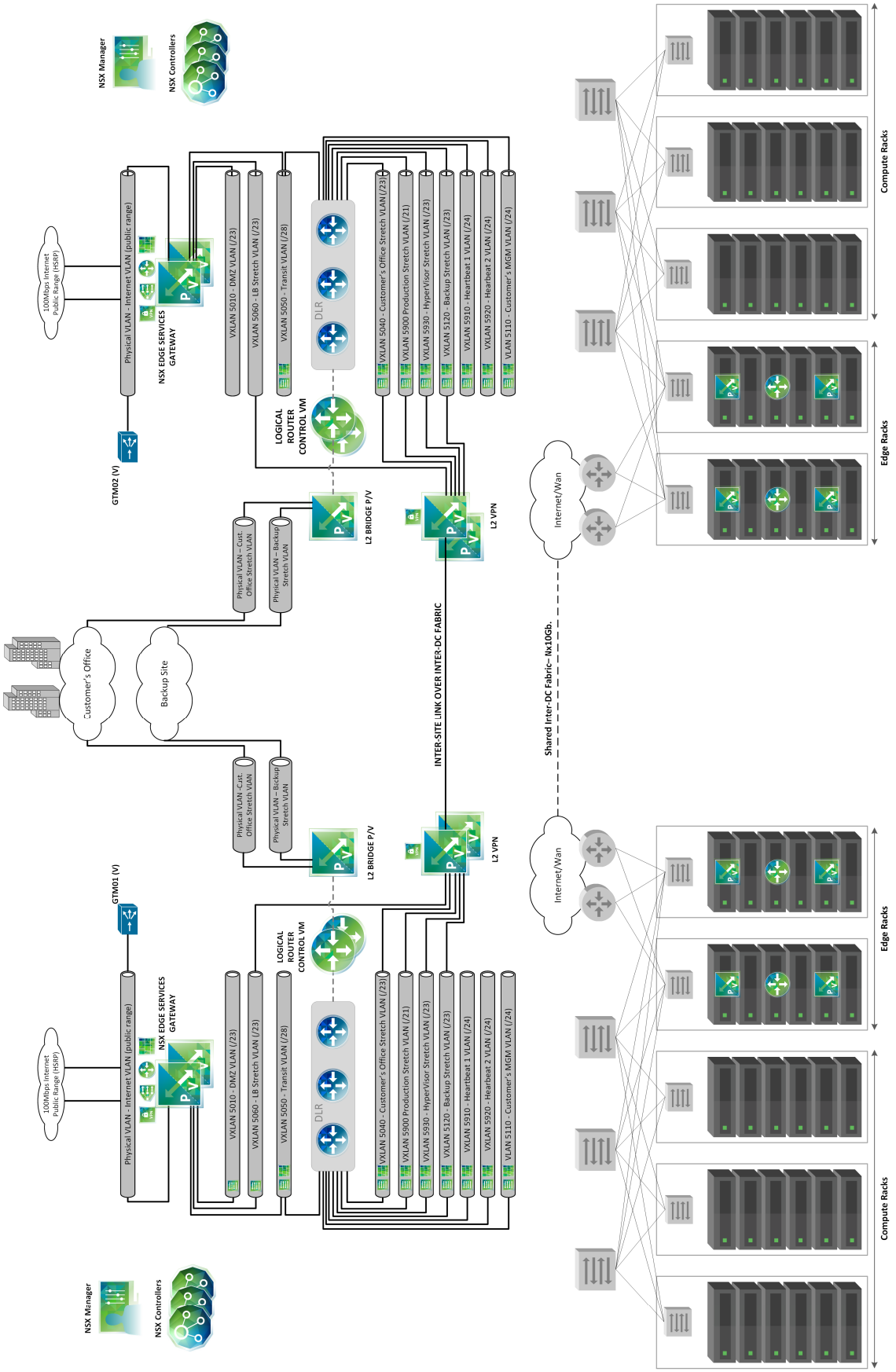

Fig 3.3. Propuesta de diseño utilizando virtualización de red

3.4. Consideraciones de diseño

En esta sección se presentan las consideraciones que se han tenido en cuenta para hacer el diseño de la solución del cliente utilizando virtualización de red con NSX a partir del diseño inicial (hecho en el capítulo 2).

En los siguientes apartados se muestra qué funcionalidades de NSX se han utilizado para substituir los distintos componentes físicos de manera que a nivel funcional se pudiera ofrecer el mismo servicio. También se incluyen las ventajas e inconvenientes o limitaciones que ofrecen los elementos en la red virtualizada.

Es importante destacar que algunas de las consideraciones de diseño hechas para la arquitectura física se conservan en este diseño como el uso de stretch VLANs (2.2.3) e interDC VLAN (2.2.4) para interconectar con los dos DCs en capa 2 y 3, así como la asignación de VLANs e IPs (2.2.1).

3.4.1. Switching lógico

Para esta propuesta de diseño de red utilizando virtualización se propone sustituir tanto los switches perimetrales, como los switches de la zona protegida utilizados en la propuesta física para interconectar los distintos equipos de red por su homologo virtual en NSX. En concreto, se utilizarán vNetwork Distributed Switch (dvSwitch, *vDS*) de VMWare junto con el soporte de VXLANs de NSX (ver Anexo D) para que las comunicaciones puedan ser transportadas sobre la red física subyacente de manera transparente.

3.4.2. NSX Edge Gateway para la zona perimetral

En la solución virtualizada con NSX los servicios de capa 3/4 de la zona perimetral se darán en la plataforma utilizando un NSX Edge Gateway en alta disponibilidad. Este equipo hará las mismas funciones que hacían tanto los firewalls perimetrales cómo los balanceadores de tráfico en la propuesta utilizando equipos físicos.

La función principal de este appliance virtual será gestionar de manera centralizada el tráfico norte-sur en la plataforma entre las redes internas desplegadas en el dominio NSX y las redes físicas externas (en este caso Internet), haciendo funciones de NAT cuándo sea necesario. También se conectarán aquí las redes LB y DMZ, dónde están servicios publicados hacia Internet para que estos puedan ser balanceados también por el equipo.

Estos equipos se encargarán del enrutamiento entre VLANs, firewall, NAT, VPNs y balanceo de carga. Para ello se activarán los siguientes servicios en el NSX Edge:

- Router: Para encaminar el tráfico entre las LANs de balanceo, DMZ, Internet y la zona de tránsito que conecta con la zona protegida de la plataforma.
- Firewall: Para segurizar las comunicaciones que pasen a través del equipo como lo hace un firewall tradicional utilizando reglas a L4: IP/s de origen, IPs de destino, protocolo y puerto.
- NAT: Entre Internet y el resto de zonas.
- VPN: VPN L3 tanto LAN to LAN utilizando IPSEC para los distintos proveedores del cliente cómo de acceso remoto utilizando en este caso VPNs basadas en SSL.
- Balanceador: Actuará de balanceador local permitiendo distribuir la carga entre los distintos miembros en las granjas de servidores, tanto en la DMZ como en cualquiera de las LANs conectadas.

A parte de estas funciones activadas de manera implícita el equipo también hará de puerta de enlace entre lo físico y lo virtual ya que interconectará la red física de Internet con el resto de redes que estarán virtualidades dentro de la plataforma.

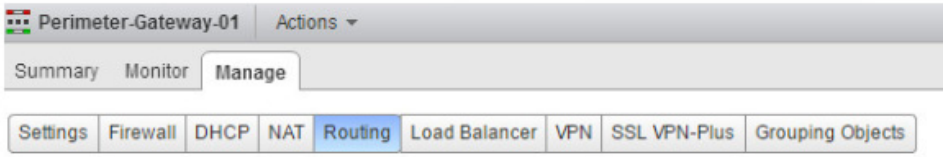

Fig 3.4. Servicios disponibles al desplegar un NSX Edge Gateway appliance

Para poder llevar a cabo todas estas funciones el despliegue de este appliance se hará utilizando la versión superior de NSX Edge Gateway llamada X-Large. Cada appliance tendrá 6 vCPUs y 8 Gb de vRAM. Se ha escogido esta opción ya que es la recomendada por el fabricante para despliegues de alto rendimiento de routing + firewall + Balanceador. El despliegue se hará siguiendo las recomendaciones de VMware en el Edge Cluster para asegurar el acceso directo a las redes físicas.

Como se ha comentado, el despliegue de los equipos se hará en alta disponibilidad. Se desplegarán dos equipos en dos ESXi distintos en modo activo/pasivo. Éstos utilizaran el primer interfaz virtual para asegurarse de que el otro equipo en la pareja está activo, sino responde el secundario, tomará todas las funciones que realizaba el equipo primario. El NSX manager se encargará de asegurar que ambos equipos están funcionando. Si por ejemplo cae el ESXi dónde está uno de ellos, lo levantará en otro de manera automática.

De acuerdo con las especificaciones del fabricante el appliance utilizando es capaz de gestionar un throughput de 9 Gbps y 1 millón de conexiones concurrentes a nivel de balanceador de carga [23] (que es la función que

necesita más recursos), muy por encima de los requerimientos del cliente. En caso de que en un futuro fuera necesario ampliar la capacidad, la plataforma permitiría de manera ágil el despliegue de un segundo par de NSX Edges usados sólo para balanceo de carga descargando así el appliance original.

Desde un punto de vista funcional, el NSX Edge teóricamente es capaz de cumplir casi todos los requerimientos del cliente. Después de testearlos se ha podido comprobar que ciertas funcionalidades son limitadas respecto a los equipos físicos.

Por lo que respecta al firewall hay algunas funcionalidades avanzadas que no están disponibles, cómo la inspección profunda de los paquetes (funciones de IDS) o la limitación a la hora de configurar NATs basados en políticas.

En cuanto al balanceador local, éste permite crear balanceos de carga para servicios TCP, UDP o HTTP/S de la misma manera que lo hacen los equipos físicos. A pesar de permitir programar balanceos más complejos mediante la utilización de scripts, estos no permiten de manera nativa la gestión de las sesiones de usuarios de Citrix u otras aplicaciones. Sera necesario delegar esta funcionalidad directamente a los frontales de la granja Citrix.

3.4.3. DLR para el enrutamiento VXLANs protegidas. Enrutamiento Lógico

Para el enrutamiento entre las redes en la zona protegida de la plataforma se utilizará un DLR (Distributed Logical Router). El DLR se encargará de interconectar todas las redes en la zona protegida y también estará conectado a la zona de tránsito para interconectar con la zona perimetral.

El DLR permite la interconexión de segmentos lógicos L2 dentro del dominio NSX. A diferencia del NSX Edge Gateway, en el DLR el enrutamiento del tráfico no se realiza en un equipo concreto, sino que se hace de manera distribuida a nivel del kernel de los hosts ESX en los que están las máquinas virtuales. Para realizar esta función son necesarios dos componentes fundamentales:

Plano de control: El plano de control del DLR se centraliza en el DLR control VM. Se trata de un appliance virtual que se encarga básicamente del soporte de protocolos de enrutamiento dinámicos (como OSPF o BGP) y subir la tabla de enrutamiento al Controller cluster para que este la distribuya a las instancias de router en los distintos ESXi. Para asegurar disponibilidad del servicio en caso de que falle este appliance (o del hosts ESX que la contiene) se propone hacer el despliegue del cliente utilizando un segundo appliance en alta disponibilidad que se activará en caso de que haya algún problema en el primario.

Plano de datos: El plano de datos del DLR son los DLR Kernel Modules (VIBs) que se encuentran instalados en los ESXi que son parte del dominio NSX. En

esta instancia distribuida de los DLRs en cada ESXi es donde se hará efectivamente el enrutamiento.

Esta arquitectura permite que el tráfico entre dos VXLANs de la infraestructura (tráfico este-oeste) se encamine a nivel de hipervisor sin necesidad de tener que subir hasta el NSX Edge Gateway o un router físico. Con esto limita el tráfico en la infraestructura física y se evita el llamado hair-pinning.

Para ilustrarlo en la Fig 3.5. se muestra la comunicación entre dos máquinas en la LAN de Producción y en la LAN de backup del cliente utilizando routing centralizado (ya sea usando un NSX Edge o un router físico). Como se puede observar el tráfico deja los racks en los que está la máquina virtual de origen para ser encaminado en los racks de comunicaciones y volviendo luego hacia la máquina de destino.

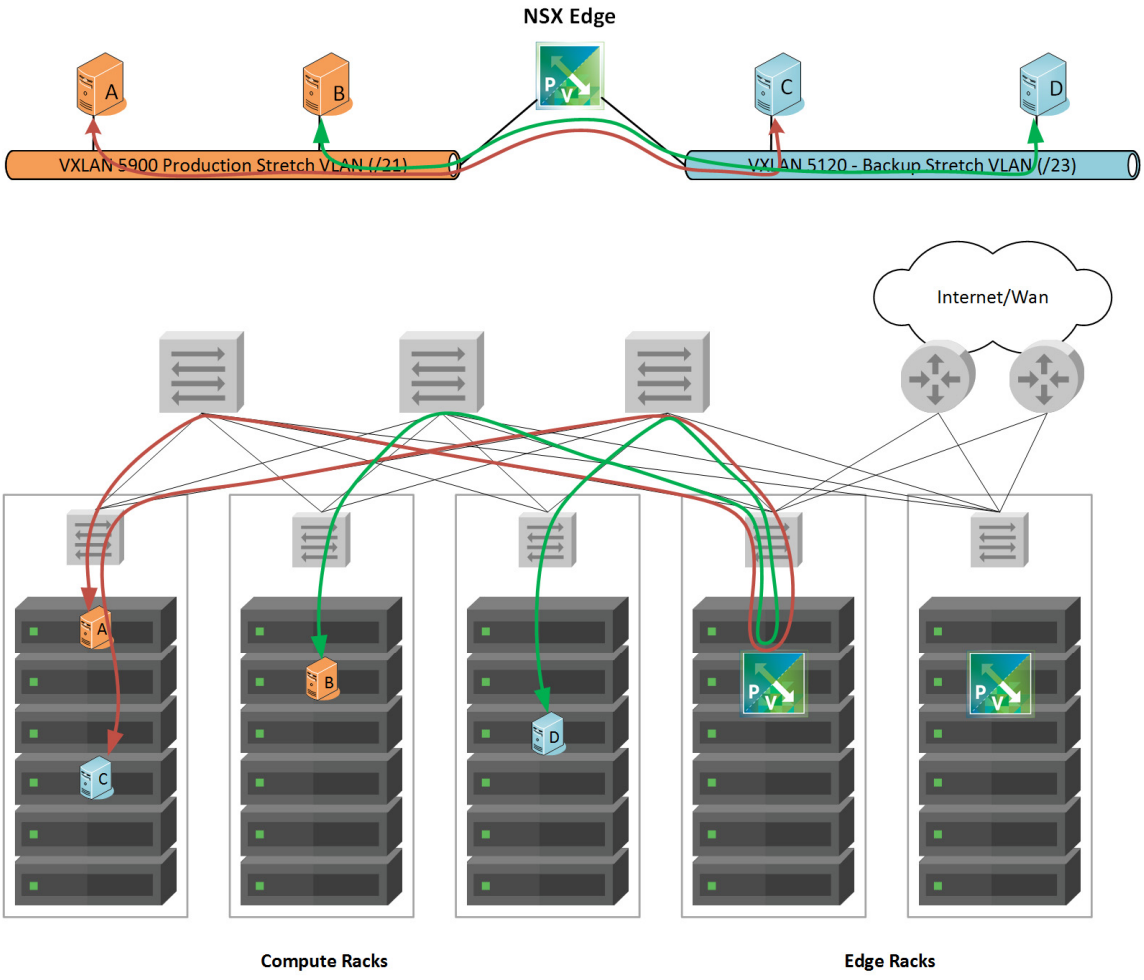

Fig 3.5. Ejemplo de tráfico entre dos VXLANs usando routing centralizado. Ejemplo de Hair-pinning.

En la Fig 3.6. se puede ver el mismo ejemplo utilizando un DLR. En este caso podemos ver que en cada ESXi tenemos una instancia del router distribuido y

además, como se puede ver en los Edge Racks, un par de DLR Control VMs para el plano de control (que no intervienen en el tráfico). Como se puede ver en este caso el camino es más óptimo gracias al enrutamiento distribuido en el kernel de los hipervisores. Para la comunicación entre los servidores B y C el tráfico va directamente entre los racks en los que se encuentras las máquinas. Para el tráfico entre A y C ni siquiera sale del rack. Si origen y destino estuvieran en el mismo servidor la comunicación ni siquiera saldría de la máquina.

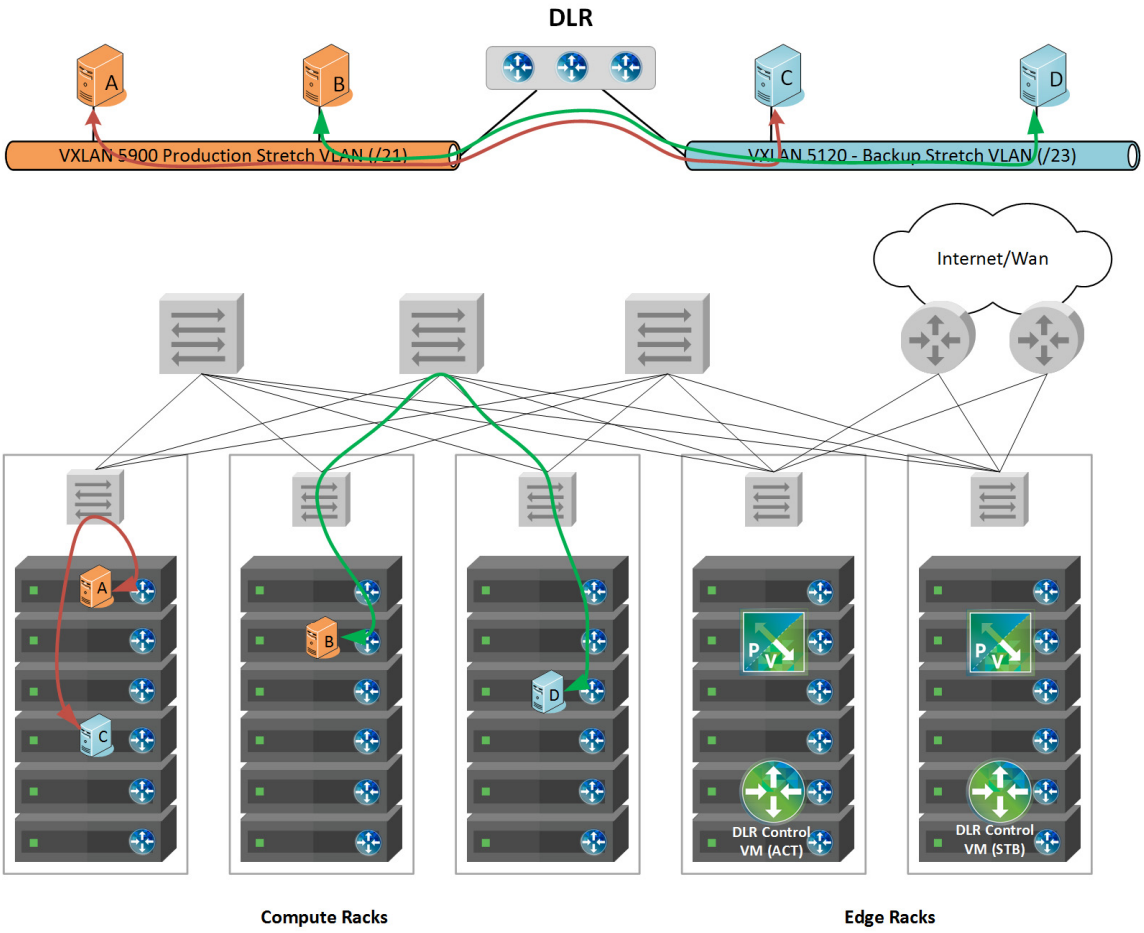

Fig 3.6. Ejemplo de tráfico entre dos VXLANs usando routing distribuido. Camino óptimo.

3.4.4. Uso DFW en la zona protegida. Firewall lógico

Para las LANs de la zona protegida, básicamente las redes conectadas a los DLR en cada localización, se ha decidido utilizar el DFW (Distributed Logical Firewall) de NSX para sustituir a los firewalls internos utilizados en la propuesta física. El objetivo de usar esta funcionalidad es proveer seguridad a la comunicación entre las redes internas. El DFW proveerá además, junto al firewall centralizado en el NSX Edge perimetral, de un segundo nivel de firewall para las comunicaciones desde el exterior hacia las redes de la zona protegida.

Al igual que el DLR, el DFW no se encuentra definido en un equipo en concreto, sino que se encuentra distribuido a nivel de kernel (usando VIBs) en los hipervisores ESXi proporcionando tasas de transferencia cercanas a la de los interfaces de red. A nivel de ESXi se crea una instancia de DFW por cada vNIC de cada una de las máquinas virtuales. Esto ayuda a mantener una buena política de seguridad ya que el filtrado se lleva a cabo lo más próximo al origen evitando que posibles amenazas se propaguen.

Las instancias DFW en cada vNIC se mueven con las máquinas virtuales cuando es necesaria su movilidad: al poner un ESXi en mantenimiento, durante la redistribución de recursos (ya sea automática – DRS [24] o manual), etc. El hecho de que las reglas de firewall tengan movilidad con las máquinas (lo que implica que no es necesario realizar ninguna intervención manual) y que estén definidas a nivel de infraestructura reduce la posibilidad de que un error humano al implementarlas comprometa la seguridad.

Implementación y reglas

Las reglas de acceso, como en el caso de los firewalls físicos, se pueden definir tanto a nivel 2 como a nivel 3/4 y son con estado. Para ello guarda una tabla de conexiones establecidas. La ventaja que ofrece utilizar el firewall integrado con la solución de virtualización de VMware es que las reglas se pueden definir utilizando los objetos (contenedores) del vCenter como origen o destino: clúster, VDS port-groups, logical switch, máquina virtual, vNICs, etc. Gracias a esto el administrador de red no necesitará saber las IPs de las máquinas e introducirlas manualmente con lo que se reducen también la posibilidad de que un fallo humano comprometa la seguridad.

Micro-segmentación

En el caso de NSX las distintas redes (VXLANs) se encuentran aisladas de las demás VXLANs. De esta manera se podrían utilizar por ejemplos las mismas redes en distintos segmentos de red y también de la red física subyacente.

En la arquitectura física propuesta la segmentación de red se realiza en el firewall que permite o deniega el tráfico entre los distintos segmentos de red. En este caso, al estar definida la instancia de DFW a nivel de vNIC se podrá realizar también micro-segmentación: las reglas no se definirán solo entre segmentos de red (VLANs en la propuesta física) sino que se podrán definir también entre servidores en que estén en la misma VXLAN. Para ilustrarlo, en la Fig.3.7. podemos ver un ejemplo concreto en algunas de las redes conectadas en el DLR del cliente, donde se implementará el DFW. En la VXLAN de producción tenemos dos pools de servidores de aplicaciones, un conjunto de servidores de copias de seguridad en la VXLAN de Backup y un par de servidores para la gestión en la VXLAN de MGM. Para facilitar la gestión los servidores están agrupados en distintos Security groups: App Servers Pool1, Apps Servers Pool2, Backup Servers y Mgm Servers. Como se puede observar en cada NIC hay una instancia del firewall distribuido en las que se implementarán las reglas mostradas en la tabla 3.1 que se ejecutan de manera secuencial.

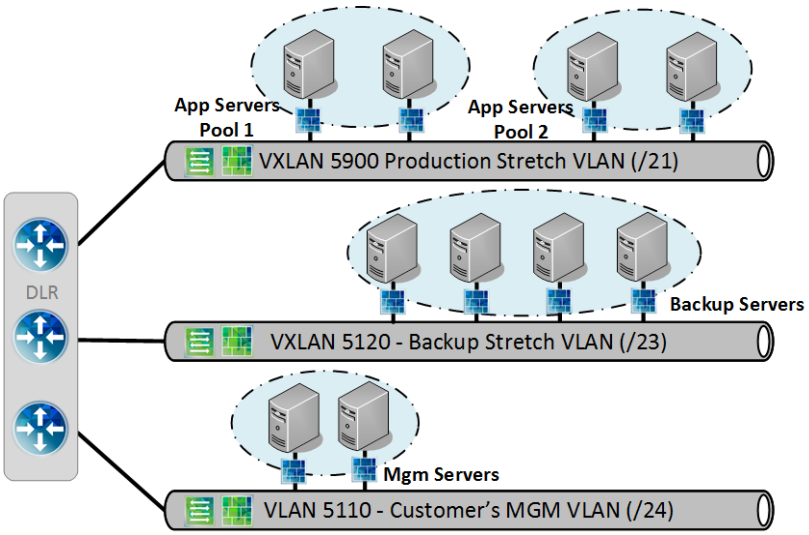

Fig 3.7. Ejemplo de uso de DFW en la arquitectura

Tabla 3.1. Ejemplo de uso de DFW. Reglas de firewall

ID	Nombre	Origen	Destino	Servicio/Puerto	Acción
1	Mgm App Srvs.	Mgm Servers	App Servers Pool1 App Servers Pool2	SSH-tcp/22	Allow
2	Mgm Bck Srvs.	Backup Servers	Mgm Servers	RDP-tcp/3389	Allow
3	Backup App Srvs.	Backup Servers	App Servers Pool1 App Servers Pool2	BCK-SRV tcp/20003	Allow
4	App1 to App1	App Servers Pool1	App Servers Pool1	icmp	Allow
5	App2 to App2	App Servers Pool2	App Servers Pool2	icmp	Allow
6	Mgm to Mgm	Mgm Servers	Mgm Servers	icmp	Allow
7	Backup to Backup	Backup Servers	Backup Servers	icmp	Allow
8	Default rule	Any	Any	Any	Deny

Las reglas 1, 2 y 3 en la tabla definen reglas entre equipos en distintos segmentos de red, de la misma manera que se haría en un firewall físico. La principal diferencia en este caso es que el enrutamiento y el filtrado se realizaran de manera distribuida en el hipervisor donde este la máquina virtual de origen.

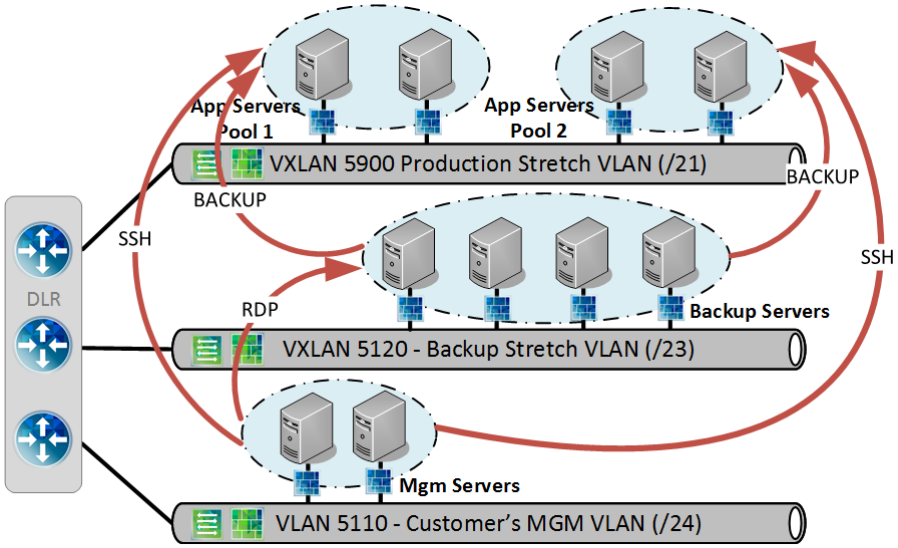

Fig 3.8. Ejemplo de uso de DFW en la arquitectura. Segmentación.

Las reglas de la 4 a la 7 definen el acceso entre equipos en la misma VXLAN (intra-VXLAN). Como hemos comentado esto no se podría hacer en una arquitectura física. La regla 8 deniega todo el tráfico no permitido en las reglas anteriores.

En la siguiente figura podemos las reglas definidas dentro de un mismo segmento de red. Todo el tráfico que no sea ICMP dentro de cada security group o cualquier otro tráfico dentro de cada VXLAN no estará permitido. Por ejemplo, no habrá visibilidad entre los servidores del App Server pool 1 y el App Server Pool 2.

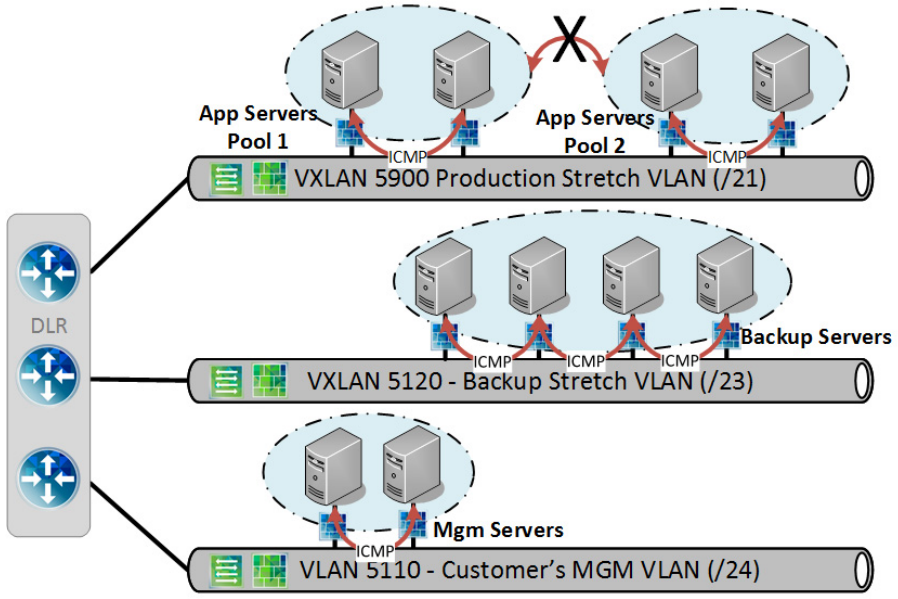

Fig 3.9. Ejemplo de uso de DFW en la arquitectura. Micro-Segmentación.

En general esta solución nos proporciona una política de seguridad mucho más granular y estricta que la arquitectura tradicional. Esto solo se puede conseguir virtualizando la red y distribuyendo este servicio.

Por lo que respecta a los inconvenientes que presenta esta solución respecto a utilizar firewalls físicos es que el DFW solo permite definir reglas en base al origen, destino y puerto. No permite definir servicios de seguridad avanzados.

3.4.5. Conexión con las oficinas del cliente y el site de backup - L2 bridge

Para las VLANs que necesitan ser extendidas a localizaciones en las que no hay NSX y que requieren la comunicación a Nivel 2 entre equipos en el entorno virtual (switch lógico – VXLAN) y equipos en el entorno físico en la localización remota (conectados a VLAN tradicional) se ha decidido utilizar la funcionalidad de L2 Bridging que ofrece la plataforma.

En concreto esto será necesario para la *Backup Strech VLAN* y la *Customer's Office Strech VLAN* ya que necesitan extenderse hacia el Site de Backup y las oficinas del cliente respectivamente en cada uno de los DCs.

- *Backup Strech VLAN* (Mapeo VLAN 120 – VXLAN 5120)
- Customer's office Stretch VLAN (Mapeo VLAN 40 – VXLAN 5040)

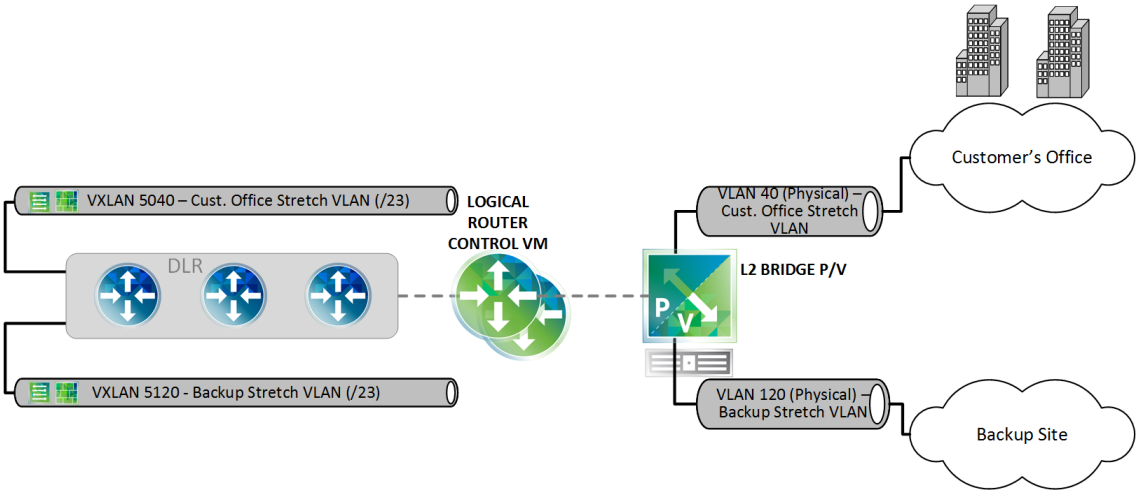

Fig 3.10. Uso de L2 bridge para VLANs extendidas a sitios remotos sin NSX

Esta funcionalidad (VXLAN-VLAN bridging) se configura a nivel de DLR y se lleva a cabo en el host ESXi en el que se encuentra la DLR Control VM. El bridging de los datos se realiza de manera completa en el Kernel del host ESXi donde está la DLR control VM. Ésta sólo determina donde se encuentra la instancia de bridging, pero no participa activamente en esta función. Sólo a nivel de configuración como se puede observar en Fig.3.11.

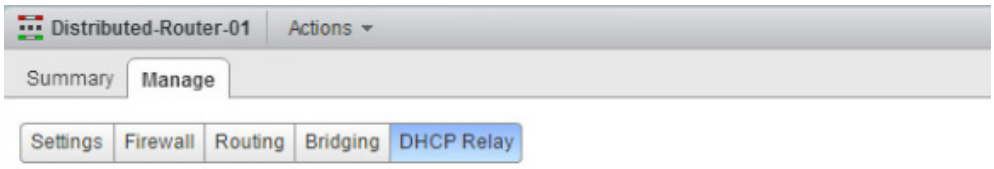

Fig 3.11. Servicios disponibles al desplegar el DLR, incluyendo el L2 Bridge.

Como el DLR Control VM se desplegará como un par de appliances virtuales en modo Activo/Pasivo para asegurar la alta disponibilidad, el bridging se llevará a cabo alternativamente en uno u otro ESXi dependiendo de cuál sea el activo en cada momento.

Siguiendo las recomendaciones del fabricante para la arquitectura, el despliegue de las DLR Control VMs se llevará a cabo en el Edge Cluster (Edge Racks o Racks de comunicaciones) con lo que solo será necesario extender las VLANs físicas a los host ESXi en este clúster y no a los hosts ESXi en el clúster dónde están las máquinas virtuales del cliente (Compute racks).

Dado el bridging se configura a nivel de DLR, ha sido necesario modificar el diseño original (y que se ha utilizado en la arquitectura física) y mover la LAN *Customer's Office* a la zona protegida donde ya usamos un DLR (antes estaba en la zona perimetral). Si no se hiciera así, como en la zona perimetral se ha utilizado un NSX Edge Gateway (ver 3.4.2), habría sido necesario desplegar un DLR solo para hacer el bridging de esta VLAN, requiriendo nuevas DLR control VMs, con el consiguiente consumo adicional de recursos y licencias.

Para estas dos LANs también se había considerado la posibilidad de utilizar VPNs L2 pero esto requería terminarla en un equipo NSX, por lo que se ha descartado. Además, los enlaces hacia las oficinas y el site de backup son dedicados, por lo que no surge la necesidad de enviar este tráfico encriptado.

3.4.6. Stretch VLANs entre los dos DCs – L2 VPN

Para las VXLANs que necesitan ser extendidas entre los dos centros de datos se ha decidido utilizar el servicio de VPNs en Capa 2 (L2 VPNs) que ofrece NSX. Este servicio crea una VPN utilizando SSL que permite transportar el tráfico en los segmentos L2 de ambas localizaciones cómo si estuvieran directamente conectados.

Para ello se utilizarán un par de NSX Edge Service Gateways en cada uno de los DCs configurados en alta disponibilidad. Estos serán independientes de los NSX Edge Gateway usado para la zona perimetral (apartado 3.4.2) y se utilizarán exclusivamente como concentradores VPN L2. Los equipos en el DC1 tomarán el rol de servidor y los equipos en el DC2 el de cliente.

A la hora de desplegar NSX Edges para ser utilizados como L2 VPNs el fabricante no ofrece información de qué capacidad tiene cada versión. Dado que esta funcionalidad no debería consumir muchos recursos (como si lo hacen

el firewall o el balanceador) se ha decidido utilizar la versión *Large* (2vCPUs/1024 MB vRAM). Si fuera necesario NSX permite pasar a la siguiente versión Quad-Large (2 vCPUs/1024MB vRAM) de manera dinámica, solo es necesario reiniciar el appliance y como estos están en alta disponibilidad esto no afectaría al servicio.

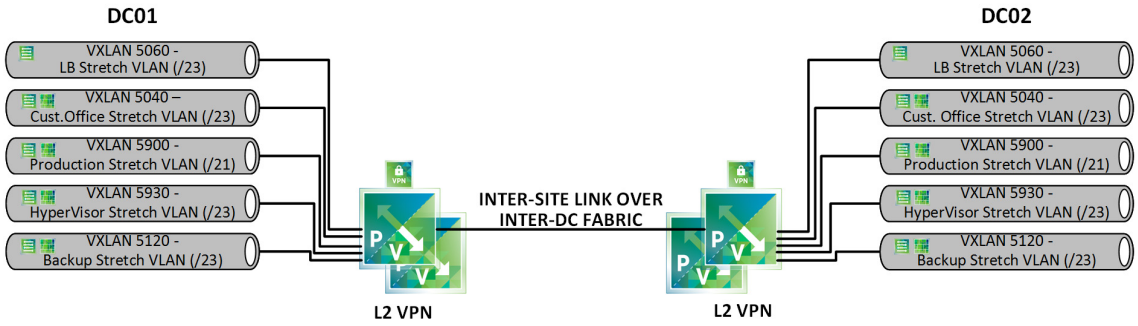

Fig 3.12. Uso de L2 VPN – Transporte de las stretch VLANS entre los dos DCs

La versión actual de NSX permite configurar un trunk para transportar múltiples VXLANs a través del túnel (en versiones anteriores estaba limitado a una VLAN/VXLAN por túnel/NSX Edge [23]). En este caso se transportarán los mismos segmentos de red que en el trunk entre DCs de la solución física incluyendo la inter-DC VLAN.

El resto de consideraciones de diseño respecto a la conectividad entre ambos DCs hechas para la arquitectura física (ver 2.2.3 y 2.2.4) se mantienen en este diseño.

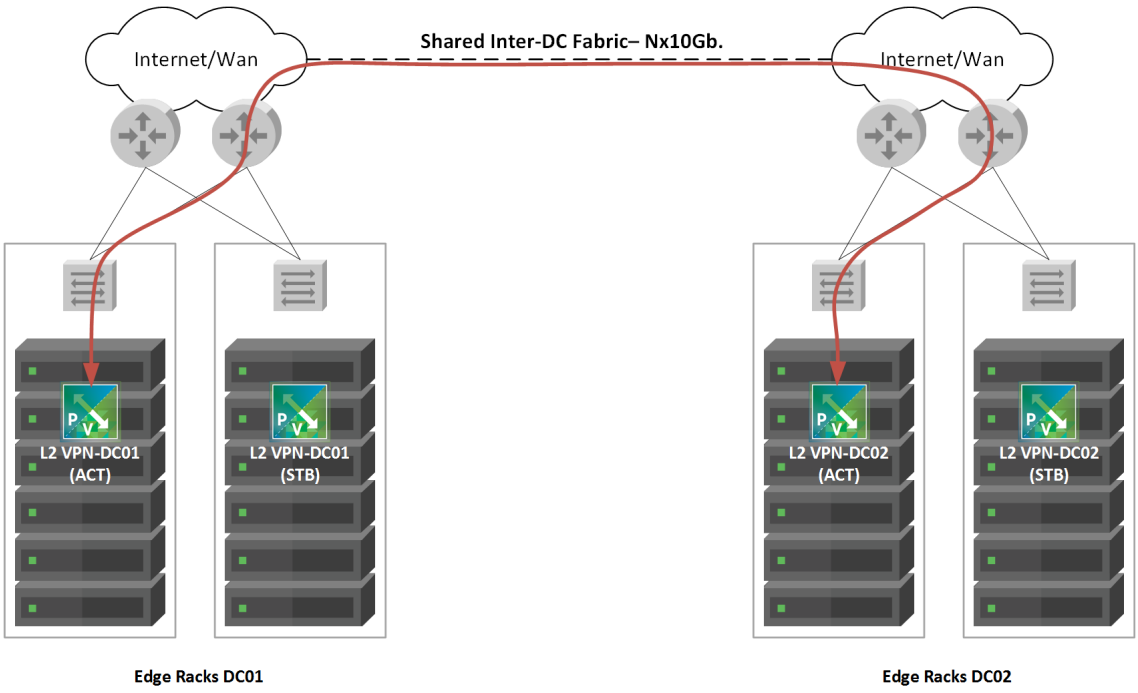

Fig 3.13. L2 VPNs – Conectividad física en la infraestructura compartida.

La comunicación entre los equipos L2 VPN se hará utilizando la infraestructura de interconexión entre los dos DCs proporcionada por el proveedor de servicios compartida en el resto de clientes de la plataforma. En este escenario (Fig. 3.13) tiene especial sentido que la comunicación esté encriptada utilizando SSL. También es importante destacar que la comunicación es independiente de la infraestructura física subyacente.

3.4.7. GTM (Global Traffic Manager) – Edición virtual.

Los GTMs (Global Traffic Managers), utilizados en la arquitectura física para balancear el tráfico entre los dos DCs y así ofrecer una solución fiable en caso de fallo grave en alguno ellos, es el único componente/funcionalidad que no se ofrece de manera nativa en la arquitectura NSX.

Es probable que en sucesivas actualizaciones este servicio, que no deja de ser un balanceo a nivel de DNS, se ofrezca de manera nativa utilizando NSX Edge Gateways al igual que actualmente se hace con el servicio balanceo local que se ha utilizado para substituir los balanceadores físicos.

Como no es posible actualmente, y esta funcionalidad es crítica para el buen funcionamiento de la plataforma, se ha decidido integrar los Big IP GTMs ofrecidos por F5 (y utilizados en la arquitectura física) en esta solución. Para mantener la consistencia con el resto de la solución se ha escogido la versión virtual Big IP V.E. (Virtual Edition) [25] en vez de un dispositivo físico.

Respecto al licenciamiento en este caso es más sencillo ya que se licencian por throughput. En este caso se ha escogido la versión Big IP V.E.25. Esta versión ofrece un throughput de 25 Mbps y hasta 1 millón de conexiones concurrentes suficiente para el servicio DNS (es un protocolo muy ligero y se esperan 30000 con/segundo). Si fuera necesario se puede actualizar de manera dinámica en el futuro actualizando la licencia y asignando más recursos a la máquina virtual cosa que no se puede hacer con los equipos físicos ya que están limitados al hardware.

Este producto es compatible con VMware y se despliega como cualquier servidor a partir de un archivo OVF (Open Virtualization Format) [26]. Como se ha decidido para la arquitectura física (ver 2.2.2) se desplegará un equipo en cada DC directamente conectado en la VLAN pública por delante de los firewalls perimetrales (En este caso el NSX Edge perimetral). Como en el caso de los NSX Edges o las DLR control VMs el despliegue se hará en el Edge Cluster. Así la VLAN pública solo tendrá que estar extendida a los hosts ESXi en este clúster y no al clúster donde están las máquinas de cliente.

CAPÍTULO 4. DISCUSIÓN

En esta sección, se pretende discutir qué ventajas e inconvenientes suponen el uso de una arquitectura tradicional basada en equipos físicos y una arquitectura basada en el uso de virtualización de red, a la hora de implementar una solución real de DC en base a los requerimientos de un cliente.

Por lo que respecta a la **solución tradicional**, utilizando equipos de red físicos, durante el proyecto hemos podido ver cómo ofrecer una solución completa en alta disponibilidad que contempla el fallo de cualquier enlace o equipo. Esto se ha hecho gracias al uso de equipos en configuración N+1 y enlaces redundantes. Además, gracias al uso de LBs globales, esta solución permite el balanceo de los servicios entre distintas localizaciones y la conmutación automática de los servicios de un DC al otro en caso de desastre. (ver Capitulo 1).

El hecho de utilizar equipos físicos dedicados para la implementación permite escogerlos en base a una gran diversidad de fabricantes (ver 2.3). Estos fabricantes, están en muchos casos especializados en funciones de red concretas (p.e. F5 en el caso de los LBs o Checkpoint en el caso de firewalls) lo que asegura un soporte dedicado con experiencia y el desarrollo específico de funcionalidades muy concretas para cada tipo de equipo. Cuando hablamos de funcionalidades concretas lo hacemos, por ejemplo, en permitir balancear aplicaciones como pueda ser Citrix, descargando a los servidores de parte de la carga, y no hacer un balanceo sólo basándose en cabeceras en Capa 3 y 4. Además, los equipos dedicados son robustos y delegan funciones específicas en hardware dedicado descargando la CPU del equipo, como sería el caso de la aceleración hardware de SSL en los LBs o el hardware dedicado utilizado en los firewalls para la encriptación de VPNs.

Desplegar una arquitectura de este tipo requiere un trabajo inicial extenso, tanto en tiempo como en trabajo a la hora de realizar el diseño y configurar los equipos. Se deben tener en cuenta factores como la redundancia de equipos y enlaces, los mecanismos o protocolos para proporcionar esta redundancia, así como la configuración de las redes lógicas que se van a desplegar a lo largo de la arquitectura. Aunque se han utilizado tecnologías como PortChannels y vPCs para reducirlo al máximo, la arquitectura en Capa 2 está basada en STP, con las limitaciones que esto supone.

Desde el punto de vista de la gestión de la capacidad, es necesario tener datos fiables de las necesidades del cliente en el momento inicial y las necesidades futuras. La principal ventaja de esto es que, a la hora de escoger los equipos, se pueden ajustar lo más posible a estos requerimientos. Si esto se hace bien, la capacidad no debería ser un elemento problemático durante la vida útil de la plataforma. El principal problema es que es poco flexible y escalable si aparecen nuevos requerimientos cuándo ya se está utilizando. Los equipos físicos tienen una capacidad limitada de escalar (puertos disponibles, capacidad del hardware…). Estos cambios pueden requerir un rediseño parcial o total de la solución y el cambio de equipos.

Desplegar nuevos servicios sobre la plataforma requiere tiempo y la reconfiguración de muchos de los equipos. Un ejemplo típico sería crear una nueva VLAN de DMZ para ofrecer un nuevo servicio desde Internet. Esto requeriría reconfigurar switches, firewalls, y LBs para que tuvieran conectividad contra ésta, reconfigurando también los distintos puertos y trunks en la plataforma. Este cambio, a priori sencillo, puede requerir del orden de días o semanas. Además, errores al configurar cualquiera de estos elementos pueden hacer caer toda la plataforma.

En caso de fallo de algún equipo, este no afectará al servicio gracias a que están todos configurados en alta disponibilidad. El único problema es que el fallo requerirá un reemplazo por parte del fabricante y la reconfiguración del nuevo equipo. En este caso es importante tener un buen sistema de copias de seguridad.

Aunque la configuración manual de los equipos es más propensa a fallos, la detección y resolución de problemas (troubleshooting) es relativamente sencilla ya que se puede revisar por separado cada uno de los dispositivos por los que pasa la comunicación que falla.

Por lo que respecta a la **solución virtualizando la red** utilizando NSX expuesta en este proyecto, también se ha conseguido una solución en alta disponibilidad tolerante a fallos. Esto ha sido posible gracias al uso de appliances virtuales en alta disponibilidad desplegados sobre distintos hipervisores y a los servicios distribuidos (switching lógico, firewalls y routers lógicos distribuidos) que se ofrecen de manera nativa en la arquitectura.

Utilizando la solución de NSX se han podido cumplir, desde un punto de vista funcional, la mayoría de los requerimientos del cliente. Se han encontrado algunas limitaciones que se explicarán más abajo y algunas funcionalidades, como el LB global para ofrecer balanceo y alta disponibilidad entre varios DCs que no están disponibles. Como esta funcionalidad era muy importante para este diseño se ha decidido integrar los mismos equipos del diseño tradicional del fabricante F5, pero utilizando su versión virtual (ver 3.4.7).

En este caso los equipos/servicios virtuales se han seleccionado entre los disponibles dentro de la arquitectura NSX para las distintas funciones (ver 3.4). En el caso de los appliances virtuales se han escogido sus distintas versiones (con distinta CPU/RAM) de acuerdo a las recomendaciones del fabricante. Esto, obviamente, simplifica el proceso de elección, pero en algunos casos las funcionalidades son limitadas. Esto no ocurre con equipos físicos de fabricantes muy especializadas en cada función de red.

El uso de esta arquitectura nos ofrece las ventajas propias de la virtualización de red usando SDN y redes de overlay, de tener servicios de red distribuidos y de utilizar appliances virtuales (SDDC) a la hora de desplegar la solución del cliente. A continuación, se enumeran estas ventajas:

- Ventajas propias de la virtualización de red SDN y uso de redes overlay:
    - Posibilidad de extender dominios en Capa 2 (LANs) en el DC de manera transparente a la arquitectura subyacente utilizando VXLANs. Esto ayuda a superar las limitaciones de las redes tradicionales en los DCs: Entre otras, se limitan los dominios broadcast, se elimina el uso de STP para la redundancia y el número VLANs ya no está limitado a 4096, lo que es especialmente interesante a nivel de ISP/proveedor de servicios de Cloud.
    - Plano de control y de datos separados, lo que facilita la gestión de ARP, tráfico unicast, broadcast y multicast en el entorno con VXLANs de manera centralizada y permite el despliegue automático de las redes virtuales.
    - No es necesario hacer ningún cambio en la topología de red subyacente. Se pueden utilizar topologías optimizadas en Capa 3 (leaf/spine). En estas topologías no hay STP (con lo que no hay enlaces infrautilizados y se mejora la estabilidad de red), se utiliza ECMP (Equal Cost MultiPath) para usar todos los enlaces disponibles de manera simultánea y son fáciles de escalar.

- Ventajas de tener servicios de red distribuidos:
    - Routing distribuido: Optimizado para el tráfico este a oeste en las redes dentro del DC.
    - Firewall distribuido e integrado con la virtualización de servidores: Permite la microsegmentación de red y que se filtre todo el tráfico que entre o salga de un servidor a nivel de vNIC aumentando la seguridad. También, al estar integrado con el entorno de virtualización permite el uso de los objetos del vCenter. Esto facilita la gestión de las reglas y dificulta cometer errores a la hora de implementarlas.

- Ventajas del uso de appliances virtuales (SDDC) como el NSX Edge de NSX o el uso de equipos virtuales como los GTM:
    - Son fáciles de escalar: Si hay problemas de capacidad se pueden aumentar los recursos (CPU, Memoria) o desplegar nuevos equipos en paralelo de forma dinámica sin necesidad de cambiar los equipos.

A pesar de las ventajas que ofrece la virtualización de red, a la hora de hacer el diseño para el cliente se han encontrado las siguientes limitaciones respecto a los requerimientos iniciales (ver capítulo 1):

- Los LBs permiten el balanceo típico en Capa 3/4 y HTTP, pero no permiten funcionalidades avanzadas como la integración de servicios como Citrix. Para conseguir este tipo de funcionalidad es necesario integrar equipos (virtuales) de otros fabricantes en la plataforma.

- El servicio de firewall es básico, especialmente en el caso del firewall distribuido. Es lógico en este caso ya que, al estar distribuido en el kernel de los hipervisores, el número de funcionalidades que se pueden añadir es limitado: No tiene por ejemplo la posibilidad de ofrecer servicios de inspección de paquetes a nivel de aplicación o servicios de IDS como requería la solución del cliente.

- Los servicios de NAT y de VPN en los NSX Edges son limitados respecto a un firewall convencional. No permiten por ejemplo el uso de NATs basados en políticas.

Por lo que respecta a la escalabilidad, la plataforma es muy flexible. Si aparecen nuevos requerimientos se puede escalar añadiendo más recursos (CPU/Memoria) a los distintos appliances virtuales o desplegando más en paralelo si es necesario para que la arquitectura pueda crecer.

Esto siempre dependerá de que haya recursos en la infraestructura física subyacente. Por esta razón quien la gestione (en este proyecto es compartida y la gestiona el proveedor de servicios) deberá hacer una gestión estricta de la capacidad. Sera necesario asegurarse que haya suficientes recursos para que las plataformas virtuales puedan crecer, añadiendo más hosts ESXi, switches Leaf o Spines, cuando sea necesario.

Utilizando una red virtualizada, desplegar nuevos servicios es rápido y sencillo. Usando el ejemplo mencionado anteriormente, si es necesario desplegar una nueva VLAN para una nueva DMZ solo será necesaria crear la VXLAN en el dominio NSX y presentarla a los equipos que requieran conectividad contra ésta. Como es transparente a la red subyacente, no será necesario reconfigurar los equipos físicos, reduciendo la posibilidad de cometer errores que afecten a toda la plataforma y reduciendo el tiempo de despliegue al orden de horas (en vez de días/meses como habíamos comentado en la solución tradicional).

Como se ha mencionado en 3.3 todos los appliances virtuales se desplegarán en alta disponibilidad (1+1). Si uno de ellos falla, por caída del servidor físico donde está desplegado, NSX lo arrancará en otro distinto. Esto evita la necesidad de sustituir el elemento físico de red como sucede en la solución tradicional. En este caso la configuración se hace de manera automática por lo que es menos propensa a fallos. Aunque esto es así, por otra parte, se puede considerar que el nivel de abstracción necesario al tener una red física subyacente y una red de overlay con servicios de red distribuidos puede complicar bastante la detección y resolución de problemas (troubleshooting).

Este proyecto se ha centrado en la comparación de las plataformas desde un punto de vista funcional. Aunque se ha intentado encontrar la información desde un punto de vista económico no se ha podido hacer una estimación exacta. Los precios recomendados por los fabricantes para los distintos equipos no se corresponden en la mayoría de los casos a los precios reales que finalmente se pagan. Estos pueden bajar hasta un 60% dependiendo del

país de compra y los acuerdos concretos de cada proveedor con el fabricante (que incluyen volumen de compra y empleados certificados). Lo mismo sucede con NSX: el fabricante no publica los precios; están también relacionados con las condiciones concretas de los proyectos y los acuerdos comerciales. Aunque se ha intentado, estos no han sido facilitados.

Los equipos físicos son muy caros, del orden de miles o decenas de miles de euros. Montar una solución de virtualización de red también lo es. La implementación de la red física subyacente y la solución NSX también lo es inicialmente. Pero el despliegue de nuevas plataformas virtuales sobre esta no aumenta significativamente el precio por lo que, montada en el entorno de un proveedor de servicios/Cloud, permite aprovechar la economía de escala, reduciendo el coste a nivel global.

Es muy importante también comparar estas dos soluciones desde el punto de vista del impacto ambiental. La solución tradicional implica el despliegue de todos los equipos mencionados, los cuales están todos encendidos consumiendo energía. Estos lo hacen tanto cuando no están realizando ninguna función activa en la plataforma (equipos en alta disponibilidad pasivos) como cuando la carga es baja: en horas de poca actividad y al principio del despliegue cuando el tráfico es bajo.

La solución de red virtualizada, en cambio, permite aprovechar las ventajas respecto al consumo energético que ofrece la virtualización desde hace años. Múltiples VMs y appliances virtuales de red se despliegan sobre cada host ESXi optimizando los recursos y evitando tener equipos infrautilizados. Además, se puede utilizar procesos que reduzcan el consumo como Host Power Management (HPM) [27] y Distributed Power Management (DPM) [28], propios de VMware. El primero es una técnica que ahorra energía poniendo parte de las VMs en un estado de consumo reducido cuando están inactivas o no necesitan ir a máxima velocidad y la segunda es una técnica que redistribuye y consolida las VMs cuando estas necesitan pocos recursos en unos hosts ESXi concretos permitiendo parar algunos de ellos, típicamente por la noche.

CONCLUSIONES

Por lo respecta a la arquitectura tradicional, se ha podido comprobar que el uso de equipos físicos de distintos fabricantes muy especializados en las distintas funciones de red permite ofrecer una solución muy ajustada a los requerimientos del cliente. Estos equipos permiten ofrecer funcionalidades avanzadas como balanceo de servidores realizando la autenticación de servicios como Citrix o funcionalidades avanzadas de seguridad como son la inspección de paquetes a nivel de aplicación o de IDS requeridas para esta solución.

Sin embargo, a la hora de implementar los mecanismos de alta disponibilidad y redundancia en la plataforma tanto a nivel de equipo como de enlace, cada fabricante tiene sus propios mecanismos, utilizando a veces protocolos propietarios, lo que implican una configuración muy específica que hace la arquitectura muy dependiente de los fabricantes escogidos.

La arquitectura virtual que se ha diseñado, montada sobre NSX de VMware que está basada en el uso de SDDC y SDN, ofrece aproximadamente las mismas funcionalidades. Esto se ha conseguido gracias al aprovechamiento de las ventajas propias de la virtualización de red usando SDN y redes de overlay, de tener servicios de red distribuidos y de utilizar appliances virtuales (SDDC).

Sin embargo, se han encontrado algunas limitaciones en la arquitectura virtual. La primera de ellas es que no existe en la arquitectura un servicio que permita el balanceo de carga global entre los distintos DCs. Por esta razón se ha optado por integrar en la arquitectura NSX el mismo balanceador de carga global usado en la arquitectura física, utilizando su versión virtual. Algunas de las funcionalidades ofrecidas por la arquitectura NSX son limitadas, como se ha podido ver en los laboratorios de pruebas que ofrece el fabricante: el balanceo de carga no permite funcionalidades avanzadas como la integración de servicios Citrix, el servicio de firewall es básico no permitiendo ni la inspección de paquetes en capa de aplicación ni servicios de IDS y los servicios de VPN y NAT son limitados. Además, hay limitaciones en cómo se ofrecen algunas funcionalidades como los bridges L2, lo que ha hecho necesarios cambios en la topología que se había propuesto inicialmente.

Por otro lado, se han podido ofrecer funcionalidades que solo están disponibles en una arquitectura virtual:

- Gracias al uso de overlay utilizando VXLANs se han podido extender redes de Capa 2 a través de la arquitectura del DC de manera transparente a la red subyacente.
- Al utilizar firewalls distribuidos integrados con la solución de virtualización de los servidores se ha podido ofrecer micro-segmentación en la red ya que todo el tráfico que entre o salga de un servidor a nivel de vNIC se filtrará. Además, al estar integrado con el entorno de virtualización de los servidores, se pueden utilizar los objetos definidos

en este, lo que evita que se cometan errores a la hora de implementar las reglas de firewall.
- Debido a que el enrutamiento se realiza de manera distribuida en los hipervisores de la solución, los flujos están optimizados para el tráfico este-oeste evitando así el *hairpinning*.

A la hora de realizar cualquier cambio en la plataforma, escalándola para ampliar el servicio existente o para desplegar uno nuevo, se ha encontrado que al hacerlo utilizando la solución tradicional ello requiere la configuración de todos los equipos. Por pequeña que sea la modificación, esta puede requerir del orden de días o semanas para ser implementada. Además, como estos equipos se tienen que configurar de manera manual, errores al configurarlos podrían hacer caer toda la plataforma. En cambio, en la plataforma virtual propuesta, al configurarse de manera centralizada utilizando un controlador y de manera transparente a la red subyacente, estos cambios se pueden hacer de manera muy ágil, pudiendo hacerse en minutos. En este caso, como la reconfiguración de la red se hace de manera automática, es menos propensa a errores que puedan afectar a la conectividad.

En cuanto a la alta disponibilidad en la solución tradicional, se ofrecerá utilizando equipos en configuración 1+1 y enlaces redundantes y en la arquitectura virtual utilizando appliances virtuales en la misma configuración o servicios distribuidos. En el caso de los equipos físicos, el fallo de cualquiera de ellos requerirá substituirlo. En la arquitectura virtual, al estar implementada sobre una arquitectura de hipervisores, si uno falla se podrá arrancar el appliance en un servidor distinto.

Debido a que la configuración de la red se hace de manera automática, la arquitectura virtual es menos propensa a fallos. En caso de que sea necesaria la detección y resolución de incidencias podría ser más complicada que en la solución tradicional debido al nivel de abstracción necesario al tener una red física subyacente y una de overlay sobre despliegan los servicios virtuales.

Desde un punto de vista económico no se ha podido hacer una estimación exacta para comparar las dos soluciones. Pero lo que sí se puede concluir es que el despliegue de nuevas plataformas virtuales sobre una red física subyacente y una plataforma de virtualización de red NSX pre-existentes, montadas en el entorno de un proveedor de servicios/Cloud, permite aprovechar la economía de escala, y no aumenta significativamente el coste a nivel global, en comparación con la solución tradicional.

Desde el punto de vista del impacto ambiental, hay claras diferencias entre las dos soluciones. La solución de red virtualizada permite aprovechar las ventajas respecto al consumo energético que ofrece la virtualización, ya que evita tener equipos infrautilizados. Además, se pueden utilizar procesos específicos dentro de la virtualización que reducen aún más el consumo, unos poniendo parte de las VMs en un estado de consumo reducido cuando están inactivas o no necesitan ir a máxima velocidad y otros redistribuyendo y consolidando las VMs cuando estas necesitan pocos recursos en unos hosts ESXi concretos.

Como conclusión final, teniendo en cuenta las ventajas e inconvenientes de cada una de las soluciones, se piensa que la solución más adecuada sería utilizar virtualización de red, pero integrando los dispositivos específicos de los distintos fabricantes cuando sean necesarias funcionalidades avanzadas. Estos dispositivos se podrían integrar en su versión virtual y así se podrían cumplir todos los requerimientos aprovechando las ventajas del uso de SDN y SDDC. También se considera que la solución virtualizada se debería montar sobre una infraestructura compartida, como en el caso de este proyecto, para aprovechar la reducción de costes asociada a la economía de escala.

El trabajo de este proyecto se podría ampliar, como se ha comentado, con una arquitectura que combine la virtualización de red con equipos de los distintos fabricantes en su versión virtual para poder ofrecer servicios avanzados.

Otro punto para trabajo futuro es el diseño en detalle de una infraestructura de NSX en un DC específico para poder implementar sobre ella las arquitecturas virtuales de sus distintos clientes.

Por ultimo sería muy interesante implementar la solución de NSX en un entorno de test para, por un lado, comprobar las distintas funcionalidades que ofrece, y por otro, desplegar encima una arquitectura como la que se ha propuesto para el cliente y validar su correcto funcionamiento. En el estadio inicial del proyecto se intentó, pero desafortunadamente, el fabricante no ofrece licencias para test o académicas que permitan hacerlo en la actualidad. Además, sería necesario un hardware bastante potente (varios servidores con requerimientos altos de CPU y RAM) para poder desplegar encima todos los componentes de la solución.

BIBLIOGRAFIA

[1] ICSA Labs, Network products certification, https://www.icsalabs.com/products

[2] F5, iRules, https://devcentral.f5.com/irules

[3] Cisco, VPC, http://www.cisco.com/c/en/us/products/collateral/switches/nexus-5000-series-switches/configuration_guide_c07-

[4] Arista, MLAG techonogy, https://eos.arista.com/mlag-basic-configuration/

[5] Gartner, ADC report, 2016, https://www.citrix.es/products/netscaler-adc/form/gartner-mq-2016-report/

[6] F5, LTM, https://f5.com/products/modules/local-traffic-manager

[7] Citrix, Netscaler, https://lac.citrix.com/products/netscaler-adc/platforms.html

[8] Common Criteria, EAL 4, http://www.commoncriteriaportal.org/products_OS.html#OS

[9] Cisco, ASA family, http://www.cisco.com/c/en/us/products/security/asa-firepower-services/index.html

[10] Cisco, Firepower, http://www.cisco.com/c/en/us/products/security/ngips/index.html

[11] Cisco, Configuring Firepower on ASA, http://www.cisco.com/c/en/us/td/docs/security/asa/quick_start/sfr/firepower-qsg.html

[12] Checkpoint, 4600 datasheet, https://www.checkpoint.com/downloads/product-related/datasheets/4600-appliance-datasheet.pdf

[13] Gatner, Enterprise Firewall report, 2016, http://blog.checkpoint.com/2016/05/27/check-point-named-a-leader-in-the-2016-gartner-magic-quadrant-for-enterprise-network-firewalls/

[14] Checkpoint, Cluster XL loadbalancing, https://sc1.checkpoint.com/documents/R76/CP_R76_ClusterXL_AdminGuide/7292.htm

[15] F5 GTM Module, https://f5.com/es/products/modules/global-traffic-manager

[16] Cisco, Catalyst 3750-x, http://www.cisco.com/c/en/us/products/switches/catalyst-3750-x-series-switches/index.html

[17] Cisco, Stack Technology, http://www.cisco.com/c/en/us/support/docs/switches/catalyst-3750-series-switches/71925-cat3750-create-switch-stks.html

[18] Arista, 7010 series, https://www.arista.com/en/products/7010-series

[19] Cisco, catalyst 4500 series, http://www.cisco.com/c/en/us/products/switches/catalyst-4500-series-switches/index.html

[20] Cisco, Nexus 5548up, http://www.cisco.com/c/en/us/products/switches/nexus-5548up-switch/index.html

[21] Arista, 7050X series, https://www.arista.com/en/products/7050x-series

[22] VMware, NSX, http://www.vmware.com/products/nsx.html

[23] VMware, NSX 6.2 documentación, https://www.vmware.com/support/pubs/nsx_pubs.html

[24] VMware, Introduction to VMware DRS and VMware HA Clusters, https://pubs.vmware.com/vsphere-50/index.jsp?topic=%2Fcom.vmware.wssdk.pg.doc_50%2FPG_Ch13_Resources.15.6.html

[25] F5, big IP virtual Edition, https://f5.com/products/deployment-methods/virtual-editions

[26] DMTF, “Open Virtualization Format”, https://www.dmtf.org/standards/ovf

[27] Wmware, Host Power Management, http://www.vmware.com/techpapers/2013/host-power-management-in-vmware-vsphere-55-10205.html

[28] VMware, Distributed Poser Management, https://www.vmware.com/techpapers/2008/vmware-distributed-power-management-concepts-and-1080.html

[29] M. Mahalingam, D. Dutt, K. Duda, “Virtual eXtensible Local Area Network (VXLAN)” RFC7348, 2014

ANEXOS

ANEXO A. ASIGNACIÓN DE IPs, VLANs Y FUNCIÓN

Nombre VLAN		Stretch (Si/No)	VLAN ID	Equipos L3	Redes DC1	Redes DC2	Función
DC1/DC2 Internet		No	499/500	ASA, GTM	62.97.100.0/24	62.67.101.0/24	Rangos públicos
DMZ		No	10	ASA, LTM	172.30.0.0/24	172.30.1.0/24	VLAN DMZ local en cada DC. Las IPs de las VIPs en el balanceador se asignarán de esta VLAN
LB Stretch		Si	60	ASA, LTM	10.251.240.0/23	10.251.241.0 /23	Los servidores DMZ/Web serán parte de esta VLAN
Customer's Network		Si	40	ASA	192.168.100.0	192.168.100.0	Red extendida hacia las oficinas del cliente
Transit		No	50	ASA, CP	172.30.14.0/28	172.30.15.0/28	Para interconectar los firewalls perimetrales (ASA) y los protegidos (checkpoint)
Production Stretch	Production	Si	900	CP	10.251.248.0 /21	10.251.251.0 /21	Cliente: Super-red de producción /21. El Gateway para el DC01 se asignará del rango 10.251.248 y el del DC02 del rango 10.251.251.x
	Virtual Hosts	Si	900		10.251.249.0/21	10.251.252.0/21	
	Reserved	Si	900		10.251.250.0/21	10.251.253.0/21	
	Global Cluster	Si	900		10.251.254.0/21	10.251.254.0/21	
	Non	Si	900		10.251.255.0/21	10.251.255.0/21	
Heartbeat 1		No	910	CP	172.30.4.0/24	172.30.5.0/24	Cliente: VLAN de Heartbeat para los servidores en clúster en la infraestructura
Heartbeat 2		No	920	CP	172.30.6.0/24	172.30.7.0/24	Cliente: 2ª VLAN de Heartbeat para los servidores en clúster en la infraestructura
Backup Stretch		Si	120	CP	10.251.232.0 /23	10.251.233.0 /23	Cliente: VLAN de backup para todos los servidores. Extendida al site de backup.
Customer's Management		No	110	CP	10.251.224.0 /24	10.251.225.0 /24	Cliente: Para Gestionar/Monitorizar los servidores.
HyperVisor		Si	930	CP	10.251.216.0/23	10.251.217.0/23	Cliente: Para gestionar los servidores virtualizados.
InterDC		Si	80	CP	172.30.12.128/26		Para enrutar VLANs en capa 3 entre los DCs.

Fig.A.1 Asignación de IPs, VLANs y función para la solución.

ANEXO B. EJEMPLO DE FUNCIONAMIENTO DE LA PLATAFORMA DE ACUERDO CON EL DISEÑO PROPUESTO

En esta sección se presenta un ejemplo de uso de la plataforma para un servicio web publicado en Internet.

B.1. Visión de conjunto

En la Fig. B.1 se puede ver una visión de conjunto de un servicio web. Los principales componentes a destacar son:

Cliente:

- Accede desde Internet

Frontend:

- VIP web: La VIP 172.30.0.10 está servida por un pool de servidores en la VLAN LB (10.241.240.101-104).
- VIP Web pública: La VIP está publicada en los firewalls con la IP 62.97.100.10.
- Los balanceadores globales mantienen la resolución del dominio www.custdomain.com apuntando de manera dinámica a las VIPs en ambos DCs: 62.97.100.10 y 62.97.101.10.

Backend:

- Servidores de aplicaciones: El servicio también está distribuido entre varios servidores, en este caso dos. El balanceo también se realiza en los Balanceadores:
    - La VIP 172.30.0.50: Pool members: 10.251.248.51 y 10.251.248.52
    - La VIP será consultada por los servidores web para servir los contenidos dinámicos en la Web, la consulta al servicio se hará de manera local y no será accesible desde Internet directamente. En este caso no es necesario NAT en el firewall.
    - Como los servidores no están detrás del balanceador será necesario activar la opción source NAT en los balanceadores para evitar problemas de enrutamiento asimétrico.
- Servidores de Bases de datos: En ellos se guardan de manera estructurada los datos necesarios para que los servidores de aplicaciones puedan servir los contenidos dinámicos.
    - En este caso se trata de un clúster de dos miembros en modo activo-pasivo. Cada servidor tiene una IP, el activo tiene la IP del clúster que es la que se atacará.
    - Miembros del clúster 10.251.248.101 y 10.251.248.102. IP del clúster 10.251.248.100

Esta es la configuración en el DC01, la configuración en el DC02 es simétrica cambiando las IPs. La Fig. B.1 muestra la configuración completa del servicio.

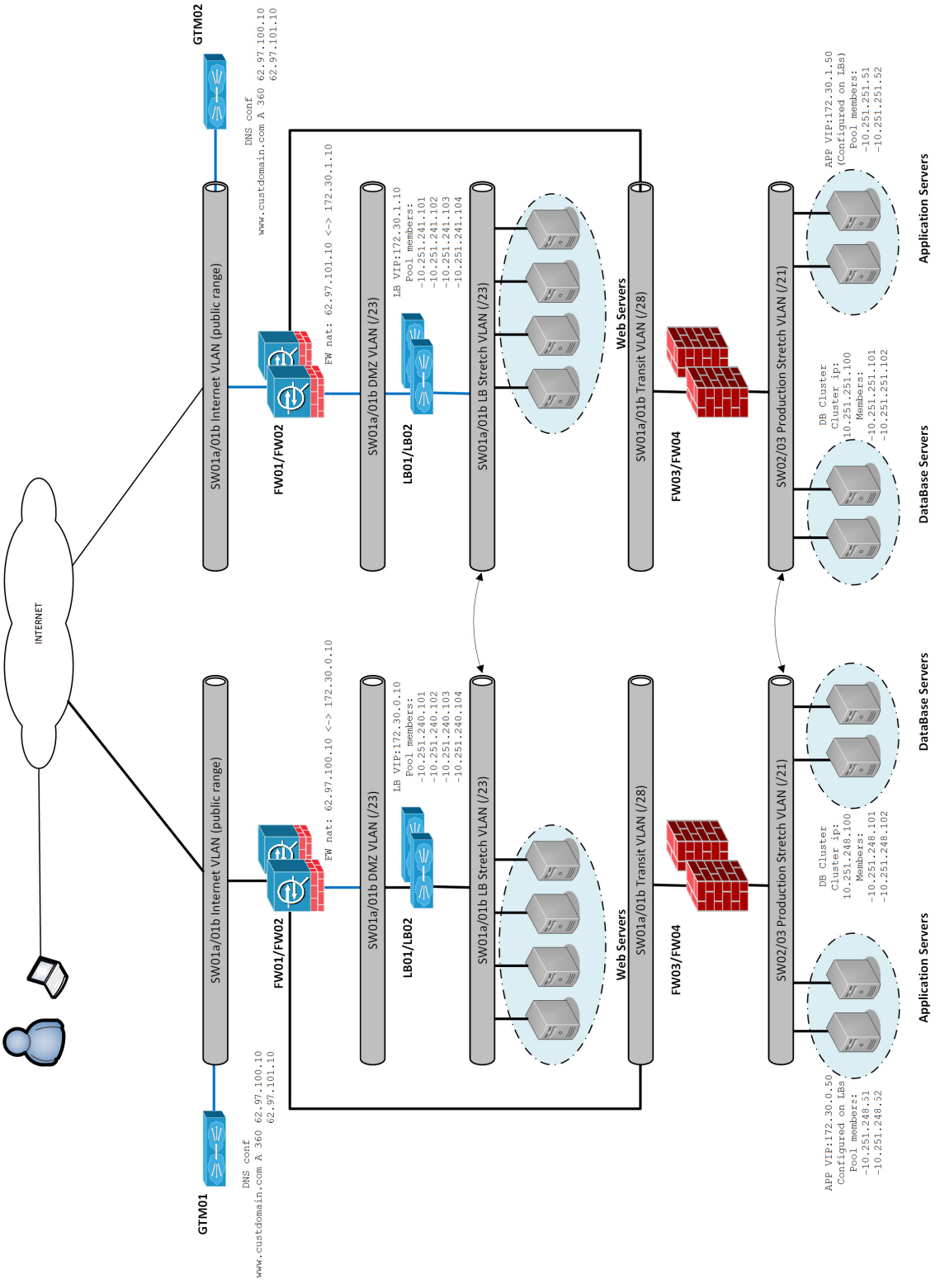

Fig. B.1. Ejemplo de uso. Visión de conjunto

B.2. Ejemplo de uso

1º Cliente realiza petición DNS:

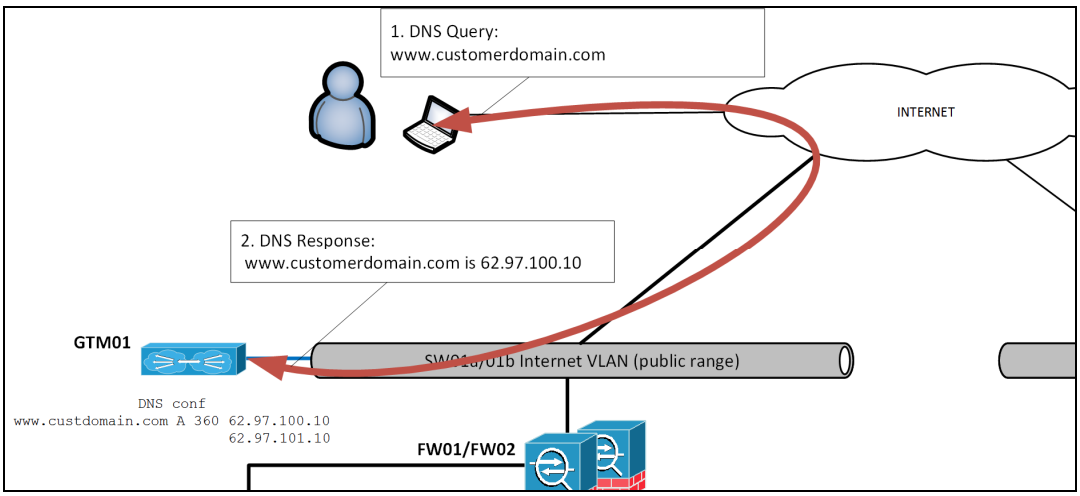

Fig. B.2. Ejemplo de uso. Paso 1.

El cliente conectado a Internet quiere acceder al dominio www.custdomain.com. Lo primero que necesita es resolver este nombre para ello envía una petición DNS.

La petición DNS se podría enviar a cualquiera de los servidores autoritativos del dominio (GTM01 en el DC01 o GTM02 en el DC02). La elección dependerá del cliente, normalmente se hace Round Robin.

Para nuestro ejemplo la petición llega al servidor GTM01, para este nombre tiene dos pool members: la IP publica de la VIP en el DC01 y la del DC02. La elección dependerá del mecanismo de balanceo; podría ser activo-backup o activo-activo y que el balanceador reparta la carga dependiendo de distintos criterios. En este caso el servidor DNS GTM01 responde con la IP 62.97.100.10 correspondiente a la VIP en el DC01.

2º El cliente realiza la petición HTTP:

Una vez resuelta la IP el cliente lanzará la petición HTTP contra la IP 62.97.100.10 que se publica desde los firewalls FW01/02 en el DC01.

Cuando la petición llega al firewall esta se natea contra la IP de la VIP en el balanceador (172.30.0.10).

El Balanceador recibe esta petición y la envía a uno de los miembros del webserver pool que se escogerá dependiendo del mecanismo de balanceo. En este caso la petición llega al servidor Web 10.251.240.102.

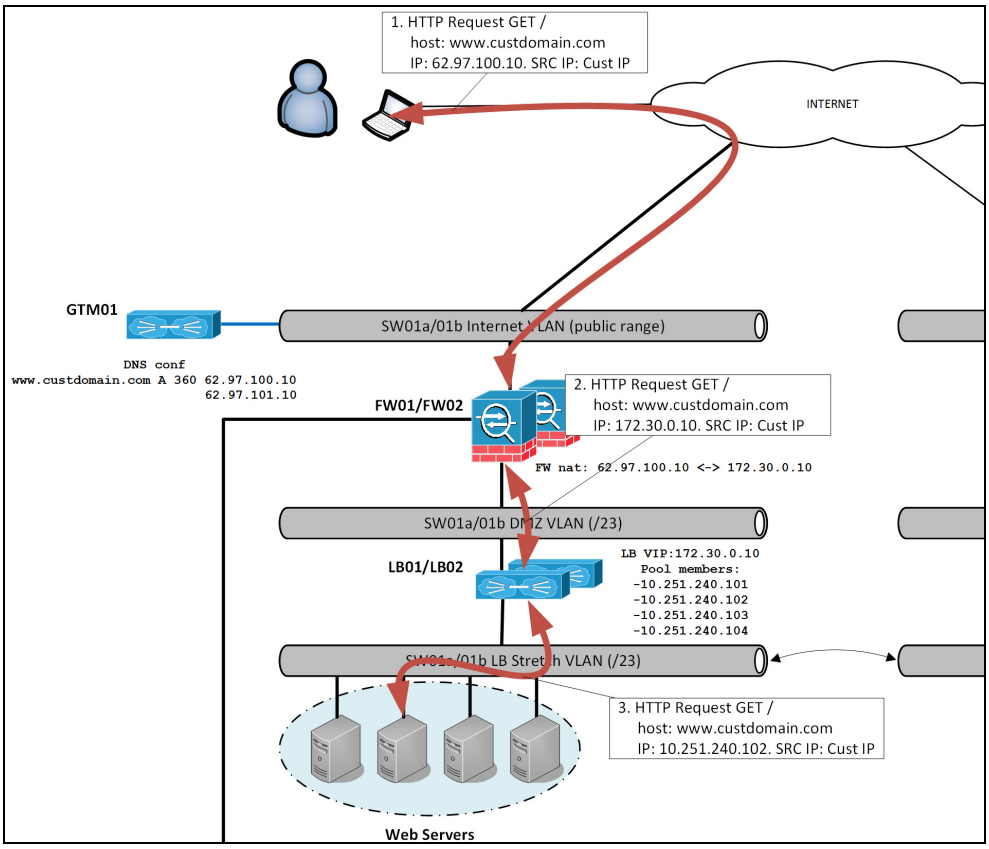

Fig. B.3. Ejemplo de uso. Paso 2.

3º El servidor web realiza consultas contra los servidores de aplicaciones (internos):

Una vez recibida la petición web por parte del cliente éste realiza una petición a los servidores internos de aplicaciones para ofrecer el contenido dinámico de la web.

Como se puede ver en la Fig. B.4 el servidor web ataca a la VIP de los servidores de aplicaciones 172.30.0.50.

Cuándo esta petición llega al balanceador, la envía a uno de miembros de la app pool, que se escogerá también dependiendo del mecanismo de balanceo. Como la VLAN donde están los servidores no está detrás del balanceador, la petición sale con la IP del balanceador (NAT de origen) para evitar problemas de enrutamiento asimétrico. En este caso la petición llega al servidor de aplicaciones 10.251.248.51.

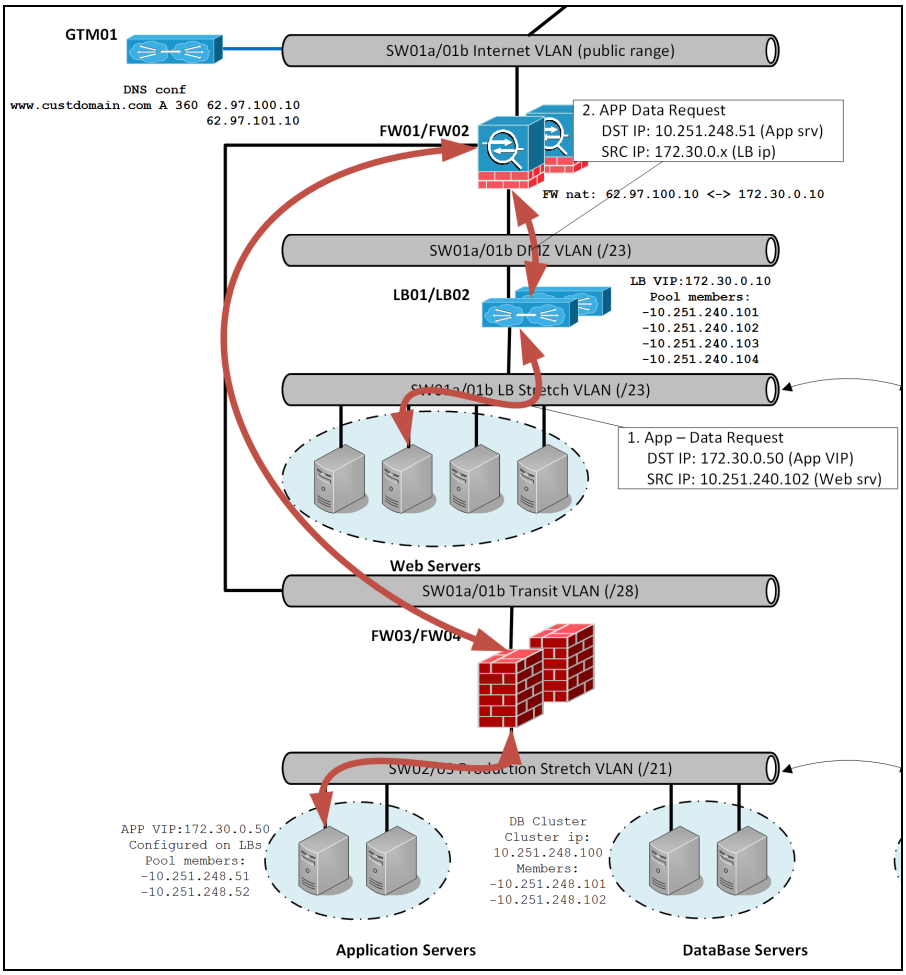

Fig. B.4. Ejemplo de uso. Paso 3.

4. Servidor de aplicaciones consulta base de datos:

Finalmente, el servidor de aplicaciones obtiene la información requerida por el cliente realizando una consulta (query) contra la base de datos. Como la BD está en clúster la consulta se hace contra la IP del clúster 10.251.248.100.

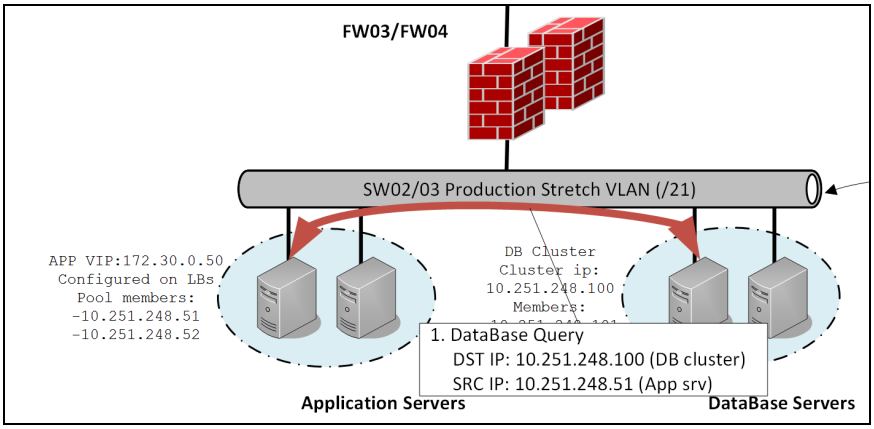

Fig. B.5. Ejemplo de uso. Paso 4.

ANEXO C. CARACTERÍSTICAS DE LOS EQUIPOS ESCOGIDOS

Tabla C.1. Principales características F5 Big IP 4000

Procesado inteligente de tráfico	Peticiones L7/s.: 425K Conexiones L4/s.: 150K Peticiones L4 HTTP/s.: 1.25M Conexiones concurrentes L4: 10M Throughput: 10 Gbps L4/L7
Procesado de paquetes SSL/TSL vía hardware	Maximo: 4,500 TPS (2K keys) 8 Gbps bulk encryption
Arquitectura	64-bit TMOS
Procesador	1 quad core Intel Xeon processor (total 8 hyperthreaded logical processing cores)
Memoria	16 GB
Disco Duro	500 GB
Puertos Gigabit Ethernet (cobre)	8 (instalados)
Puertos 10 Gigabit Fibra (SFP+):	2 (opcionales)
Fuente de alimentación	1 x 400W Incluida Dual o DC (opcional)

Tabla C.2. Principales características F5 Big IP 2000s

Procesado inteligente de tráfico	Peticiones L7/s.: 212K Peticiones L4/s.: 75K Peticiones L4 HTTP/s.: 550K Conexiones concurrentes L4: 5M Throughput: 5 Gbps L4/L7
Arquitectura	64-bit TMOS
Procesador	1 dual core Intel Xeon processor (total 4 hyperthreaded logical processing cores)
Memoria	8 GB
Disco Duro	500 GB
Puertos Gigabit Ethernet (cobre)	8 (instalados)
Puertos 10 Gigabit Fibra (SFP+):	2 (opcionales)
Fuente de alimentación	1 x 400W Incluida Dual o DC (opcional)

Tabla C.3. Principales características ASA 5515-X

Stateful inspection throughput (max)	1.2 Gbps
Stateful inspection throughput (multiprotocol)	600 Mbps
IPsec site-to-site VPN peers	250-
High availability	Active/Active - Active/Standby
Puertos Gigabit Ethernet (cobre)	6 (Integrados)
Expansión	6 puertos Gigabit ethernet (cobre o SPF)
Memoria	8 Gb
Flash	8 Gb

Tabla C.4. Principales características Checkpoint 4600 Series

Firewall Throughput (Testing)	9 Gbps.
Firewall Throughput (Real-world)	3,4 Gbps.
Sesiones concurrentes	1.400.000
High availability	Active/Active - Active/Standby
Puertos 10/100/1000Base-T	8 (integrados)
Expansión	4 puertos Gigabit Ethernet (cobre o SPF)
Disco Duro	250 GB
Memoria	4 Gb

ANEXO D. NSX-V. COMPONENTES Y FUNCIONAMIENTO

Con el objetivo de facilitar la comprensión de la solución NSX en la siguiente sección se ofrece un resumen de los distintos elementos y funciones que conforman la solución. Para más detalles se recomienda consultar documentación NSX [23].

D.1. Introducción

La arquitectura de NSX propone virtualizar la red en el datacenter (DC) de la misma manera que actualmente lo está haciendo con los servidores. NSX de manera programática crea, borra, crea imágenes y restaura redes virtuales basadas en software. Todo esto se realiza de manera transparente a la red subyacente y sin necesidad de hacer cambios en ella.

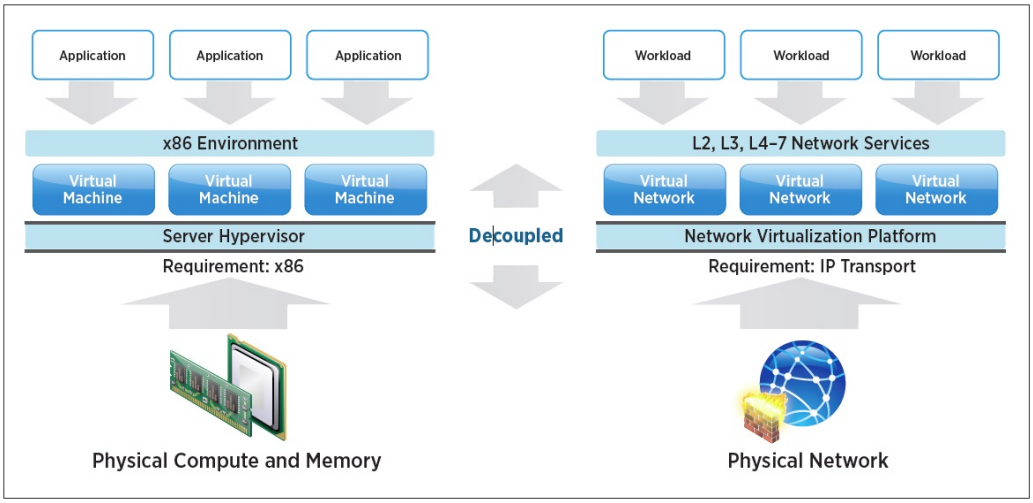

Fig. D.1. Analogía entre redes virtuales y servidores virtuales

El despliegue NSX se basa en tres componentes principales: El plano de gestión, el plano de control y el plano de datos. Fig. D.2.

Plano de control

Esta función se lleva a cabo en el **NSX Controller** que permite la creación de VXLANs y la programación de elementos como el Distributed Logical Routing (DLR).

Se trata de un elemento solo de control, los datos no pasan a través de él. Los controller nodes se despliegan en clúster de un número impar de miembros

(mínimo 3) para asegurar la alta disponibilidad y la ausencia de problemas de Split-brain si alguno de los nodos cae o queda aislado.

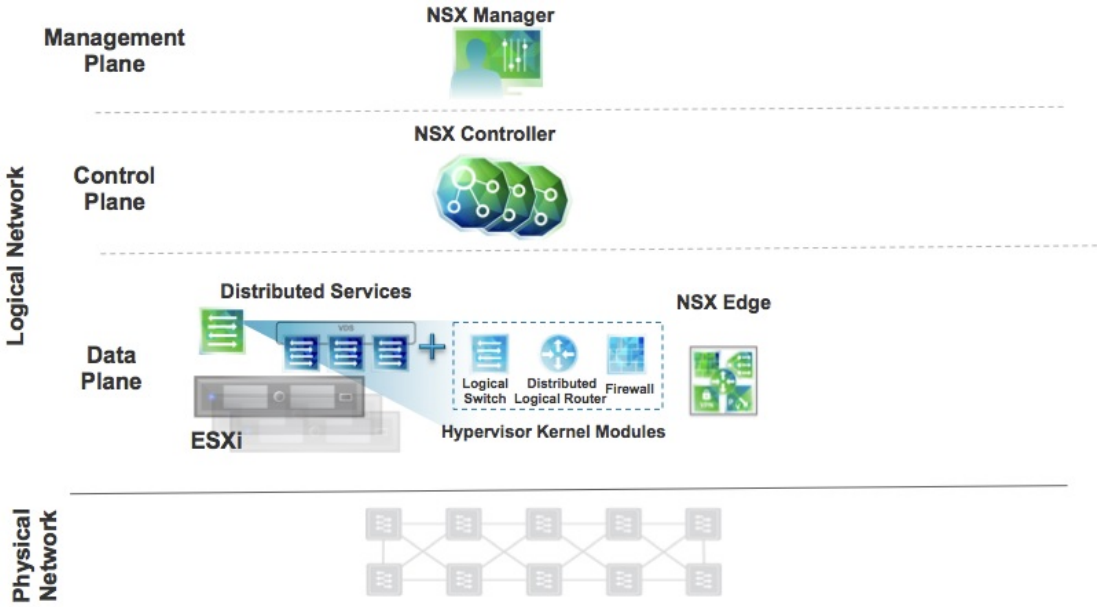

Fig. D.2. Componentes NSX

Plano de datos

El componente principal del plano de datos es el **NSX vSwitch** que está basado en el VSphere Virtual Distributed switch (VDS) usado en Vmware con componentes adicionales para poder habilitar servicios añadidos. Estos incluyen módulos de kernel (VIBs) que se ejecutan en el kernel del hipervisor para proveer servicios como routing distribuido, firewall distribuido o habilitar el uso de VXLANs.

El NSX VDS abstrae la red física y proporciona switching en el hipervisor independiente de elementos físicos como VLANs. Sus principales ventajas son:

- Permiten el soporte de una red de overlay con el uso del protocolo VXLAN y la gestión y configuración centralizada de la red. Este overlay permite.
    - Crear una red overlay L2 sobre las redes IP existentes, en la infraestructura física existente sin la necesidad de rediseñar estas redes.
    - Provisión flexible de comunicaciones (tanto este-oeste como norte-sur) manteniendo el aislamiento entre las redes.
    - La red de overlay es transparente para aplicaciones y máquinas virtuales, que operan normalmente como si estuvieran conectadas a redes L2 físicas.
- Facilitan escalado de hipervisores de manera masiva.

- Permiten el uso de características que facilitan la gestión, monitorización y troubleshooting en la red virtual. Algunos ejemplos son: Port Mirroring, NetFlow/IPFIX, QoS, LACP…

El plano de datos también está compuesto por equipos que hacen de gateway permitiendo comunicar entre el espacio de red lógico (VXLAN) y el espacio de red físico (VLAN). Esta funcionalidad se puede realizar a nivel 2 (bridging) o a nivel 3 (routing).

Plano de gestión

En esta arquitectura este rol lo asume el **NSX Manager**. Este proporciona un punto único de configuración para toda la arquitectura utilizando su interfaz de usuario web. También puede integrarse con orquestadores de plataformas de cloud utilizando REST APIs.

Servicios funcionales

Sobre esta arquitectura NSX reproduce de manera fiel servicios de red y seguridad en software:

- Switching: Permite la extensión de segmentos de red L2 / Subredes IP de manera transparente a la arquitectura de red subyacente.
- Routing: Permite enrutar tráfico entre subredes IPs de manera lógica sin necesidad de salir a un router físico. Esto se hace en el kernel de los hipervisores permitiendo un camino óptimo para el tráfico dentro de la infraestructura virtual (este-oeste).
  De la misma manera el componente llamado NSX Edge proporciona un punto centralizado para la integración con la red física gestionando la comunicación con las redes externas (Norte-Sur).
- Firewall: Se hace de dos maneras también: a nivel de kernel, de manera distribuida por Virtual NIC (Network Interface) o también a nivel del NSX Edge (normalmente para la comunicación con redes externas).
- Balanceador de carga: Permite crear balanceos a L4-L7 de red con capacidad también usarlos de terminación SSL (SSL ofload).
- VPN: Servicios L2 Y L3.
- Conectividad con redes físicas: Las funciones Gateway L2/l3 permiten la integración elementos en la red física.

D.2. Componentes funcionales

D.2.1. NSX Manager

El NSX mánager es un virtual appliance que se encarga de la configuración de switches lógicos y de conectar las máquinas virtuales (VMs) a estos switches lógicos. Aunque no es parte de este proyecto también es el punto de entrada donde se configuran las APIs de NSX que permiten gestionar la plataforma de manera orquestada utilizando servicios de Cloud.

Se despliega en entornos VMware con una relación 1:1 por cada vCenter Server trabajando fuertemente ligados. Como se despliega como VM, puede utilizar las funciones de alta disponibilidad si cae para que VMware la levante en otro equipo físico. La caída del NSX Manager no afecta a las redes virtuales ya desplegadas.

El NSX manager se encarga del despliegue del Controller Cluster y de la reparación de los hosts ESXi para trabajar con NSX instalando una serie de módulos VSphere installation Bundles (VIBs) para habilitar VXLANs, instancias de firewall y router distribuidas y la comunicación con el plano de control. También es responsable de desplegar los NSX Edge Gateway appliances y los servicios asociados de los que se hablará a continuación. Por ultimo también se encarga de la seguridad en el plano de control habilitando certificados para autenticar la comunicación entre los distintos elementos.

D.2.2. Controller Clúster

Es el elemento en el plano de control de la solución que se encarga de manejar los módulos de routing y switching en los hipervisores. El uso del Controller Cluster para manejar VXLANs elimina la necesidad de multicast en la red física subyacente o la inundación de paquetes ARP en broadcast dentro de cada segmento de red. Se recomienda consultar la documentación de NSX para ver los distintos modos de replicación VXLAN y como suprime el uso de ARP

Como se ha comentado para evitar temas de Split-Brain el Controller Cluster se despliega en un número impar de nodos (3 o más). Para aumentar la escalabilidad de la solución estos nodos se reparten la gestión de las distintas instancias de routing y switching montadas en la arquitectura. Para cada uno de estos roles hay un máster (que se elige por mayoría, al ser impares), que se encarga de repartir estas instancias (o trozos – slices). Si uno de los nodos cae, los otros se reparten la gestión de las instancias que gestionaba.

D.2.3. Uso de VXLANs

El uso de redes de overlay se está volviendo muy popular por su capacidad de separar la conectividad en el espacio lógico de la infraestructura de red física. A nivel lógico se pueden configurar un conjunto de servicios de red mientras que la red física se convierte únicamente en una capa de transporte.Esta

separación permite resolver muchos de los retos que supone utilizar una red física en el datacenter:

- Permite el despliegue de aplicaciones de manera ágil y rápida. En las redes tradicionales por ejemplo crear y extender una nueva VLAN por toda la red puede llevar días.
- Permite la movilidad de servidores virtuales en el DC. En los DCs tradicionales esto requeriría extender el dominio L2 (VLANs) a través del DC afectando a la escalabilidad y potencialmente a la resiliencia de la arquitectura.
- Limitaciones de STP (Spanning-Tree Protocol). STP es el mecanismo que se utiliza tradicionalmente para evitar bucles y proporcionar resiliencia en capa 2. Sin embargo, STP deja múltiples enlaces sin uso cosa que a gran escala supone un enorme costo para puertos y enlaces que no se van a utilizar. El uso de redes de overlay permite la extensión de dominios de capa 2 sobre redes en capa 3 que no tienen las limitaciones de spanning-tree y permiten por ejemplo utilizar todos los enlaces al mismo tiempo.
- Infraestructuras multi-cliente a gran escala. En las redes tradicionales el número máximo de redes aisladas que se pueden tener es de 4096 (correspondiente al número máximo de VLANs soportadas). No suele ser un problema en redes empresariales, pero se está convirtiendo en una importante limitación en proveedores de servicios en la nube.

Debido a que es apoyado por los principales fabricantes, VXLAN (Virtual Extensible LAN) se ha convertido en un estándar de facto en las redes overlay. El uso de VXLANs es clave para construir redes lógicas sin las limitaciones que se encuentran en las redes tradicionales. Este estándar se encuentra definido en el RFC7348 [29].

VXLAN es un protocolo se encarga de encapsular los paquetes ethernet generados por los distintos servidores (ya sean físicos o virtuales) conectados al mismo segmento de red L2 y transportarlos de manera transparente a través de la red L2/L3 subyacente.

En la Fig. D.3. podemos ver formato del paquete VXLAN. Como se puede observar el estándar propone encapsular el paquete L2 generado por el origen añadiendo nuevas cabeceras (VXLAN, UDP, IP y Ethernet) de manera que este pueda ser transportado entre los VTEPs (VXLAN Tunnel End Points).

A destacar de la cabecera:

- VXLAN NI (VNI): Se trata de un identificador de 24 bits utilizado para identificar el segmento lógico de red (segmento VXLAN). Esto permite crear 2^24 (más de 16 millones) de redes.
- Puerto de origen UDP: Se utiliza un hash de las cabeceras L2/L3/L4 originales. El objetivo es asegurar el balanceo del tráfico entre caminos del mismo coste en la infraestructura subyacente.

- Puerto de destino UDP: Es fijo, en el RFC es el 4789, NSX utiliza de manera propietaria el puerto 8472.
- IP Origen/IP Destino. Corresponden a las IPs de los equipos que realizan la encapsulación VXLAN (los VTEPs). En el caso de NSX este rol lo toman los hosts VMware ESXi normalmente, aunque también equipos de otros fabricantes que sean compatibles con la solución.

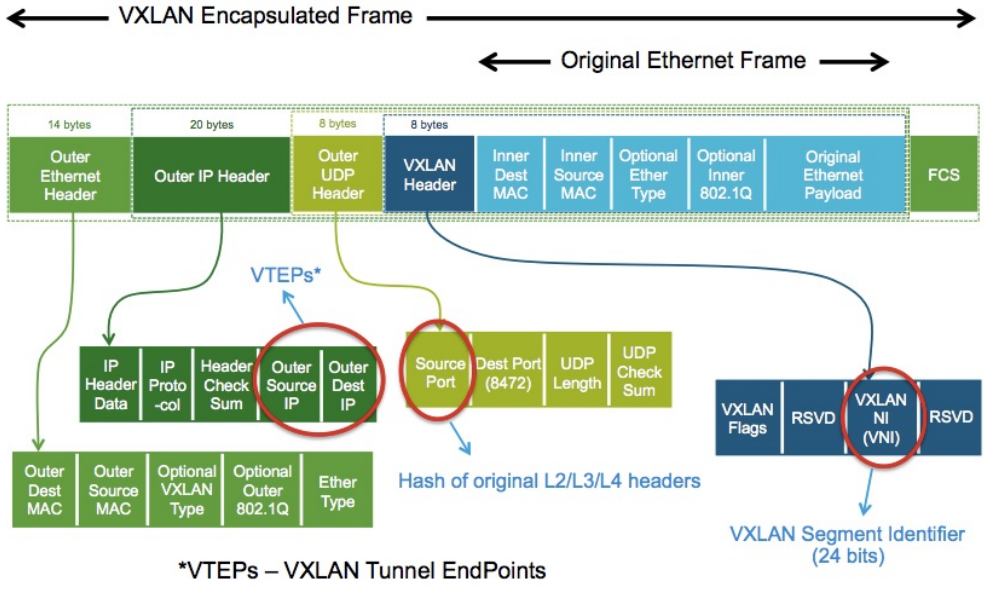

Fig. D.3. Encapsulación VXLAN – Formato del paquete

Como estas cabeceras incrementan el tamaño del paquete IP resultante se recomienda cambiar la MTU en la infraestructura subyacente al menos a 1600 bytes.

D.2.4. Hipervisores ESXi con VDS

El VDS (Virtual Distributed Switch), que se encuentra distribuido en el kernel de los distintos hosts ESXi es el componente principal para la integración de NSX.

NSX requiere la instalación de una serie de módulos a nivel de Kernel de los hipervisores (VIBs. VMware Installation Bundles) para habilitar las funcionalidades necesarias en NSX: switching, routing y firewall distribuido, así como encapsulación VXLAN.

A continuación, se describe cómo interactúan con el plano de control y de gestión de NSX en la arquitectura.

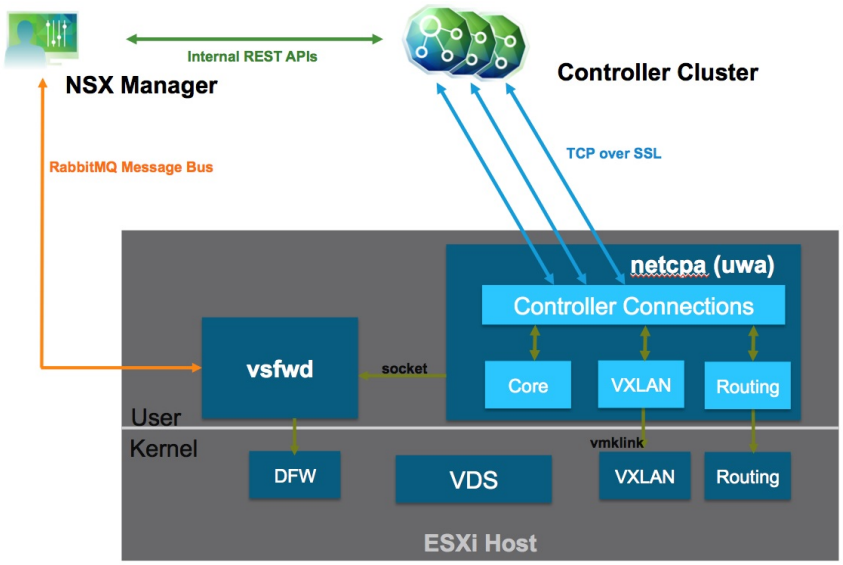

Fig. D.4. Interacción entre host ESXi y el plano de gestión y control de NSX

- RabbitMQ Message Bus: Comunicación entre el NSX Manager y el host ESXi. Se utiliza para enviar las reglas para ser instaladas el firewall distribuido, las IPs de los controller nodes, claves privadas y certificados para autenticar la comunicación entre hosts y controladores y para crear/borrar instancias de router distribuido.

- User World Agent (UWA). Este proceso establece múltiples sesiones TCP sobre SSL con los controladores. Esto crea un canal de control por el comparten las tablas MAC, ARP y de VTEPs. El objetivo principal es controlar dónde están conectados los servidores virtuales en la red lógica distribuida desplegada.

D.2.5. NSX EDGE service gateway

Se trata de un componente básico en la arquitectura NSX sobre VMware ya que permite proveer múltiples servicios en ella. Los servicios disponibles son:

- Enrutamiento y NAT: Proporciona routing centralizado entre las redes lógicas en el dominio NSX y la infraestructura de red física externa. Soporte múltiples protocolos de encaminamiento dinámicos como OSPF, iBGP o eBGP, así como rutas estáticas. También puede hacer NAT de origen o destino.
- Firewall: Mientras que el DFW se utiliza normalmente para securizar la comunicación entre equipos conectados a la red lógica, el NSX Edge se utiliza normalmente para securizar las comunicaciones entre la red física y la red lógica (tráfico norte-sur).
- Balanceador de carga: Proporciona balanceo de carga L4-L7.
- VPNs L2 y L3: Las VPNs L2 se utilizan normalmente para interconectar dos dominios de capa 2 en localizaciones REMOTAS. Las VPNs L3

pueden ser IPSEC site-to-site o de acceso remoto conectando utilizando SSL.
- DHCP y DNS: Puede funcionar de DNS relay y de servidor o relay DHCP.

Este componente se despliega en forma de appliance virtual cosa que permite un despliegue rápido y escalable. Este despliegue se puede hacer utilizando distintas configuraciones dependiendo de las funcionalidades a habilitar.

Tabla D.1. Configuraciones para el despliegue de NSX Edge Gateway

	vCPUs	vRAM (MB)	Notas
Compact	1	512	
Large	2	1024	
Quad-Large	4	1024	Indicado para firewall de alto rendimiento
X-Large	8	8192	Indicado para Firewall + Balanceador + Router de alto rendimiento

Se puede cambiar la configuración después del despliegue inicial, pero esto requiere un pequeño corte de servicio.

Por lo que respecta a la alta disponibilidad y resiliencia los NSX Edges se pueden desplegar como un par de equipos en modo Activo/Pasivo. El NSX manager despliega los dos equipos en hosts ESXi distintos (utilizando las reglas de anti-afinidad de VMware). Los dos equipos se envían mensajes de keepalive y si el activo no está disponible el pasivo toma este rol.

D.2.6. Zona de transporte

La zona de transporte es el conjunto de ESXi que pueden comunicarse utilizando la infraestructura física subyacente utilizando VXLANs. Cada uno de ellos se convierte VTEP y se encarga de encapsular/desencapsular estos paquetes. Los paquetes encapsulados se transmiten utilizando la infraestructura subyacente.

La zona de transporte puede extender a lo largo de uno o múltiples clústers de ESXis. Por ejemplo, el clúster donde están las VMs y el clúster donde se despliegan los servicios de red como los Edge Gateways.

D.3. Servicios Funcionales

A diferencia de lo que sucede con las redes tradicionales, la arquitectura NSX no depende de equipos independientes para realizar las funciones de red (routers, switches, firewalls…). Estas funciones se realizan utilizando todos los componentes explicados en la sección anterior de manera conjunta.

Dependiendo de la función, se realizan de manera centralizada, distribuida o ambas a la vez. De manera centralizada normalmente se utilizan los NSX Edges. De manera distribuida el Controller Cluster crea reglas de procesamiento que se instalan en los distintos hosts ESXi para que simulen una porción del servicio lógico que representan.

D.3.1. Switching lógico

Gracias a la separación entre la red física subyacente y la red lógica que crea NSX, esta permite crear segmentos de red lógicos aislados de manera dinámica sobre la infraestructura, permitiendo que varias VMs en distintos ESXi en cualquier parte del DC se vean como si estuvieran físicamente conectadas a la misma VLAN.

Para conseguir esto, NSX utiliza la tecnología de overlay VXLAN para extender los switches lógicos sobre la infraestructura física. Cada switch lógico está definido por un VXLAN ID. En cada uno de estos switches lógicos se pueden conectar tanto máquinas virtuales como físicas. Para integrar estas máquinas físicas son necesarios servicios de bridge L2 que se explicaran en apartados posteriores.

En esta arquitectura, para cada máquina conectada a una VXLAN, el Controller Cluster mantiene de manera centralizada la información relevante para que otras máquinas se puedan conectar con ella estén donde estén. Esta información la comparten con los VTEPs para que sepan donde enviar los paquetes encapsulados a través de la red de transporte física.

Tabla D.2. Ejemplo información mantenida por el Controller Cluster

VNI	IP1	MAC	VTEP
5001	IP1	MAC1	IP-ESXi1
5001	IP2	MAC2	IP-ESXI2

Por ejemplo, en la tabla anterior, podemos ver la información mantenida para dos VMs en la misma VXLAN (5001) que están en dos ESXi distintos. Si VM1 envía un paquete hacia la VM2 el ESXi1 encapsulará este paquete, se pondrá como IP de origen en la cabecera externa (IP-ESXi1) y la del ESXi2 como destino (IP-ESXI2). Este paquete será transportado por la red física hasta el

ESX2 que lo desencapsulará y lo enviara a la VM02 como si las dos máquinas estuvieran físicamente conectadas.

Para ver como se rellenan estas tablas, como se gestionan los distintos de tipos de tráfico (Multicast, Unicast…) se recomienda consultar la documentación de NSX.

D.3.2. Enrutamiento lógico

La funcionalidad de enrutamiento lógico de NSX permite la interconexión de dispositivos (tanto virtuales como físicos) desplegados en las distintas redes lógicas en capa 2. Esto es gracias a la separación entre infraestructura física y la red lógica desplegada por la virtualización de red.

El enrutamiento lógico se puede utilizar para dos propósitos fundamentales:

- Interconectar dispositivos (físicos o lógicos) conectados a dos redes lógicas distintas (comunicación este-oeste).
- Interconectar dispositivos conectados a una red lógica con dispositivos en la red física externa (comunicación norte-sur)

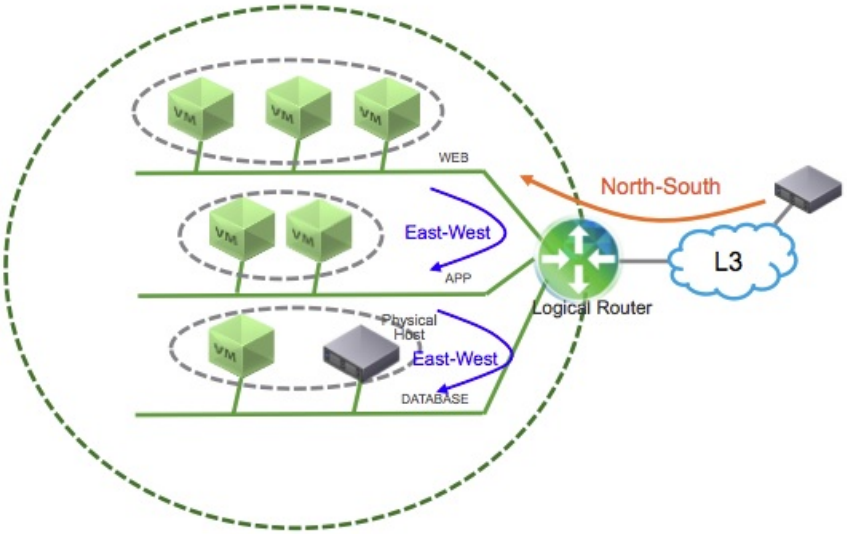

Fig. D.5. Ejemplo de enrutamiento lógico

El enrutamiento lógico se consigue normalmente utilizando dos funcionalidades en NSX: enrutamiento Centralizado y enrutamiento distribuido

D.3.2.1. Enrutamiento Centralizado

El enrutamiento centralizado permite la comunicación entre redes lógicas y las entre redes lógicas y las redes de capa 3 externas.

En NSX, el NSX Edge Services Gateway proporciona el routing tradicional centralizado. A parte de esto, como se ha comentado ya, proporciona otros servicios como DHCP, NAT, Firewall, Balanceador y VPNs.

Este enrutamiento centralizado puede utilizarse tanto para el tráfico este-oeste como para el tráfico norte-sur. Sin embargo, en el caso del tráfico este-oeste, el tráfico no está optimizado ya que, en el peor de los casos, para comunicar dos máquinas en dos redes distintas en el mismo host, ESXi tiene que salir hasta el rack de comunicaciones donde está el NSX Edge. Lo que es conocido como *hair-pinning*.

D.3.2.2. Enrutamiento Distribuido. Distributed Logical Router (DLR)

El enrutamiento distribuido permite la comunicación entre las redes lógicas dentro de la infraestructura.

Para ello el enrutamiento se encuentra distribuido a nivel de hipervisor. Gracias a esto se evita el efecto del hair-pinning comentado anteriormente. En cada uno de los hipervisores se instala información específica de los flujos (recibida desde los NSX Controllers) asegurado un camino de comunicación directa, aunque las máquinas pertenezcan a redes lógicas distintas.

Esta funcionalidad la proporciona un elemento lógico llamado Distributed Logical Router (DLR) que tiene dos componentes fundamentales:

- Plano de control: El plano de control del DLR se centraliza en el DLR Control VM. Se trata de una appliance virtual que se encarga básicamente del soporte de protocolos de enrutamiento dinámicos (como OSPF o BGP), intercambiar actualizaciones de encaminamiento con el siguiente salto L3 y de la comunicación con el NSX Manager y el NSX Controller Cluster. Para asegurar la continuidad del servicio en caso de fallo de esta máquina (o del host ESX que la contiene) se puede desplegar como un par de appliances en modo activo-pasivo de la misma manera que se puede hacer con los NSX Edge Gateways.
- Plano de datos: El plano de datos del DLR son los DLR Kernel Modules (VIBs) que se encuentran instalados en los hipervisores que son parte del dominio NSX. Estos módulos a nivel de kernel tienen una RIB (Routing Information Base) que van actualizando de manera dinámica desde el Controller Cluster. Los módulos de kernel vienen equipados con interfaces lógicas llamadas LIF (Logical Interfaces) conectadas a los distintos switches lógicos. Cada LIF tiene una dirección IP representando el Gateway para el segmento L2 conectado.

En la siguiente figura se puede ver la interacción de los distintos elementos para conseguirlo:

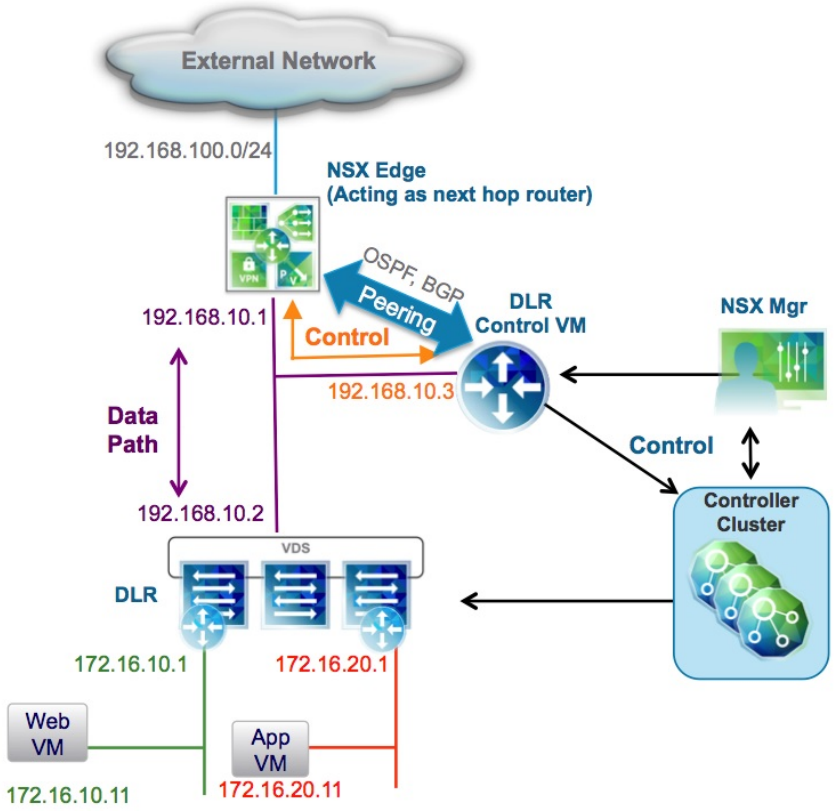

Fig. D.6. Componentes del enrutamiento lógico distribuido (utilizando DLR)

Como podemos ver en la figura, desde el NSX manager se despliegan las DLR control VMs que son las que se encargan directamente de la gestión de los protocolos de encaminamiento (por ejemplo, OSPF o BGP) intercambiando mensajes de enrutamiento con los vecinos (peering) para aprender las rutas.

Las DLR Control VMs envían las rutas aprendidas al Controller Cluster y éste se encarga de distribuirlas a los módulos DLR en el kernel de los distintos ESXi. Estos módulos son los que se encargan de la gestión del tráfico de acuerdo a estas rutas.

Para asegurar la distribución de la carga cada miembro del Controller Cluster se encarga de la gestión de cada una de las instancias de routing desplegadas en la arquitectura.

En la siguiente figura (Fig.D.7) se puede ver el camino que sigue la comunicación para dos máquinas VM1 y VM2 que se encuentran en dos segmentos de red distintos. Como se puede observar el enrutamiento se realiza en la instancia de DLR en el host1 dónde está VM1.

También es importante destacar que el enrutamiento siempre es local, realizándose en el host ESXi donde se encuentra la VM de origen. Así, para la comunicación desde la VM2 a la VM1 el enrutamiento se realizará en el host2.

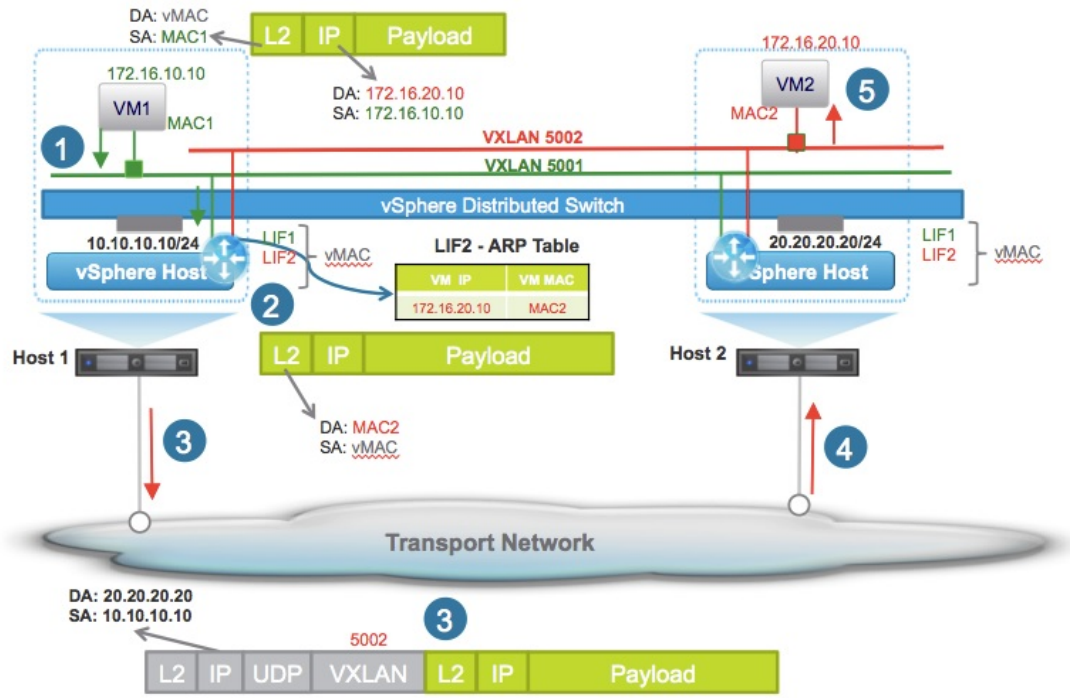

Fig. D.7. Ejemplo de comunicación utilizando el enrutamiento lógico distribuido

D.3.3. Firewall lógico

En lo que respecta a la seguridad, NSX soporta dos funcionalidades. Por un lado, proporciona servicios de firewall lógico para la protección entre las distintas redes en la plataforma y por otro permite la inserción de servicios de seguridad de otros fabricantes dentro de la plataforma.

El firewall lógico, al igual que en el caso del routing, lo proporciona de dos maneras: un servicio de firewall centralizado utilizando NSX Edges y otro servicio de firewall distribuido a nivel del kernel de los ESXi, llamado NSX Distributed Firewall (DFW).

D.3.3.1. Firewall centralizado

El servicio de firewall lógico centralizado se ofrece utilizando appliances NSX Edge. Este equipo virtual permite la segmentación de las redes conectadas filtrando el tráfico que pasa a través de él de la misma manera que lo haría un firewall tradicional.

Como se puede hacer en cualquier firewall tradicional, el firewall centralizado permite definir reglas de acceso L3/L4 utilizando IP/s de origen, IPs/s de destino y servicio/puerto.

Dado que el NSX Edge hace de Gateway hacia la red física éste está especialmente indicado para securizar los flujos norte-sur en la plataforma y desde la red física hacia la red lógica. Igualmente, también se puede utilizar

para securizar las comunicaciones entre las redes lógicas conectadas al appliance.

D.3.3.2. Firewall distribuido. NSX Distributed Firewall (DFW)

El NSX Distributed Firewall (DFW) es una funcionalidad de NSX que proporciona firewall en capas L2-L4 con estados. Este se ejecuta en el kernel de cada uno de los ESXi hosts creando una instancia de DFW por cada VNIC de cada máquina virtual. Si una VM no necesita DFW se puede excluir. Por defecto se excluyen las VMs de la infraestructura como el NSX Manager, los NSX Controllers o los NSX Edge Gateways desplegados.

Las reglas de DFW se pueden definir de dos maneras: creando reglas en capa 2 (Ethernet) o reglas en capa 3/4 (IP/Puerto).

- Reglas L2: Se crean definiendo MACs de origen, MACs de destino y servicios en capa dos. Estas son prioritarias respecto las reglas L3/L4.
- Reglas L3/L4. Se crean utilizando IPs de origen, IPs de destino y puertos TCP/UDP.

El objetivo fundamental del DFW es proteger el tráfico entre las máquinas conectadas a la infraestructura ya sean virtuales o físicas (tráfico este-oeste). Como se definen a nivel de vNIC también se utilizan para definir las reglas entre las máquinas y la red física. El DFW se utiliza de manera complementaria al firewall ofrecido por el NSX Edge Gateway. El EG se utiliza típicamente para proteger el tráfico norte-sur siendo la puerta de entrada al SDDC.

Como opera a nivel de vNIC las VM siempre esta protegidas independientemente de cómo estén conectadas a la red lógica.

La arquitectura del DFW está basada en tres componentes fundamentales:

- vCenter Server: Es el plano de gestión para el servicio. Las reglas se crean aquí, esto permite utilizar los contenedores del vCenter (Cluster, VDS-portgroup, Logical Switch, VM, vNIC, Resource pool…) como origen o destino de las reglas. También se pueden utilizar IPs o puertos directamente de manera alternativa.
- NSX Manager: Es el plano de control para el servicio. Recibe las reglas desde el vCenter y las guarda de manera centralizada. El NSX Manager se encarga de instalar estas reglas en los distintos hosts ESXi.
- ESXi host: Es el plano de datos del servicio: Recibe las reglas desde el NSX Manager y las ejecuta en el kernel para el tráfico de las VMs. Se ejecutan a nivel de vNIC así que si una VM (A) se quiere conectar con otra (B) en otro ESXi éstas se ejecutaran cuando el tráfico sale de A en el primer ESXi y cuando lleguen a B en el otro ESXi.

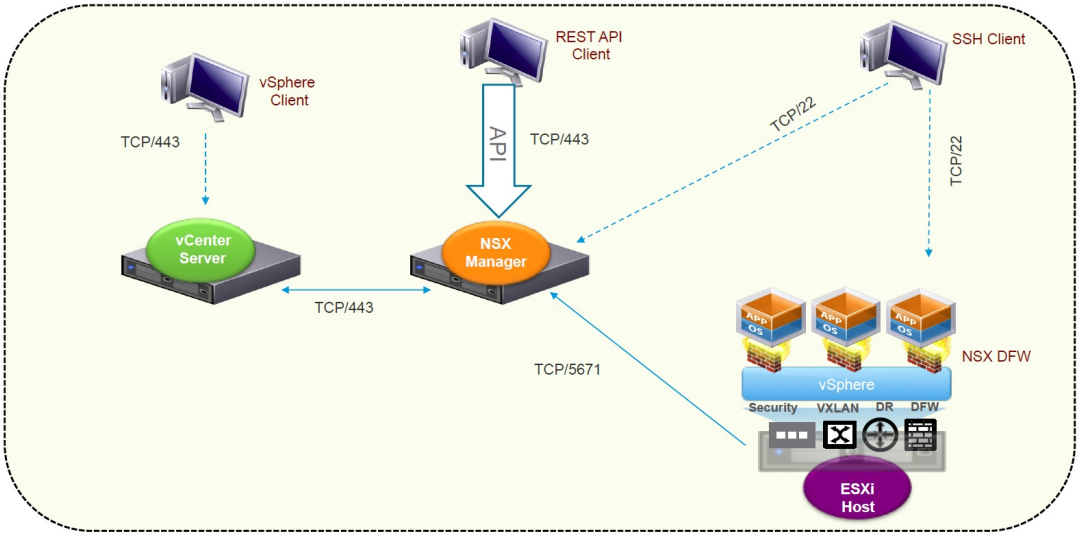

Fig. D.8. Arquitectura DFW

D.3.3.2.1. Componentes

El DFW se activa durante la preparación del ESXi. En esta operación se carga un Kernel VIB llamado VSIP (VMware Internetworking Service Insertion Platform). El VSIP se encarga de proteger el tráfico de las VMs (es el motor del DFW en sí).

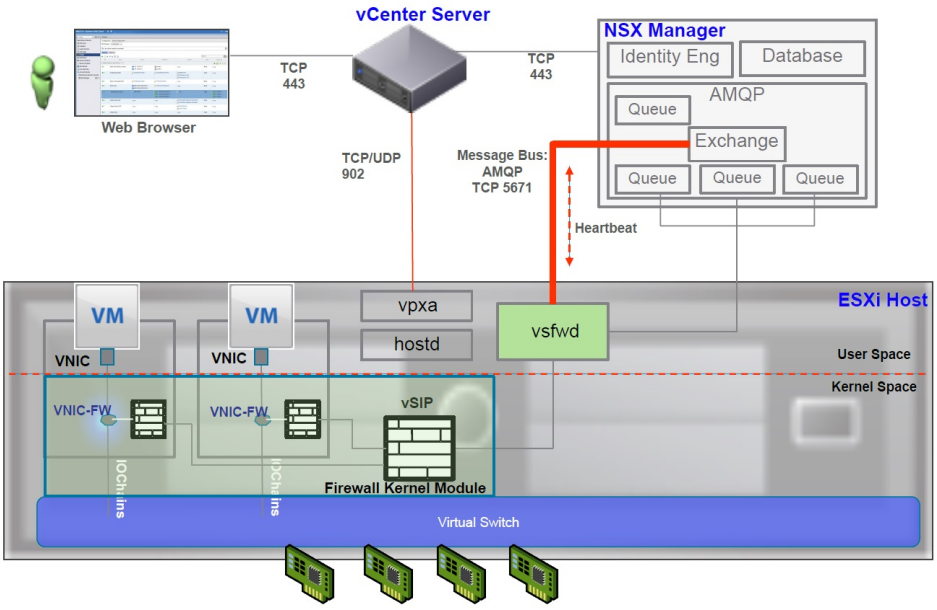

Fig. D.9. DFW - Componentes

A parte, se crea una instancia del DFW por vNIC colocada entre la VM y el virtual switch al que está conectada, por lo que cualquier tráfico que entre o salga de la maquina pasara por él.

Por último, el ESXi corre un demonio llamado vsfwd que se encarga de tareas como recibir las reglas desde el NSX manager y enviar estadísticas y logs de vuelta hacia el NSX Manager.

Además de las funciones de firewall el VSIP se encarga del anti-spoofing y de la redirección del tráfico (o un tipo particular de tráfico) a equipos de terceros para ofrecer funcionalidades de firewall avanzadas. Por ejemplo, funcionalidades de inspección profunda de paquetes (en capa 4-7) o de IDS.

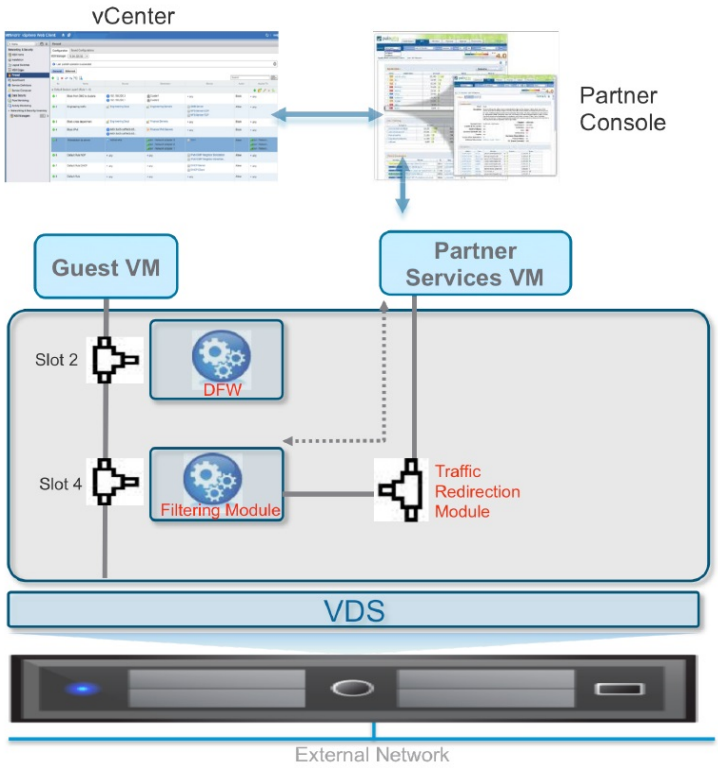

Fig. D.10. DFW - Redirección a equipos de terceros para funcionalidades avanzadas de firewall

Cada instancia de DFW a nivel de vNIC contiene dos tablas:

- Tabla de reglas: Se leen de manera secuencial, cuando el tráfico que coincide con la regla es denegado o permitido (si es permitido se guarda en la tabla de conexiones), sino pasa a la siguiente regla hasta llegar al final. Normalmente la última regla deniega todo lo que no se ha permitido anteriormente.
- Tabla de conexiones: Con las conexiones ya establecidas (y previamente permitidas).

Cuando las VMs se mueven de un ESX a otro (ya sea de manera automática o manual) las dos tablas se mueven con ella, con lo que no hay corte del tráfico durante la migración.

D.3.3.2.2. Microsegmentación

En un entorno clásico de firewall se protege el tráfico entre distintos segmentos de red. Las máquinas conectadas a un mismo segmento (VLAN, VXLAN…) no están protegidas ya que tienen visibilidad directa sin pasar por el firewall.

En este caso, ya que la instancia de firewall DFW está definida a nivel de vNIC cualquier tráfico enviado o recibido por la VM es sistemáticamente inspeccionado. Esto permite proteger el tráfico entre dos VMs en distintos segmentos de red (como lo hace un firewall tradicional) pero también si se encuentran en el mismo segmento de red. Esto introduce un nuevo concepto, la microsegmentación, que proporciona una capa de seguridad adicional que solo se puede conseguir al virtualizar esta funcionalidad de red.

D.3.4. Balanceo de carga lógico

El balanceo lógico es otra de las funcionalidades que se pueden habilitar en los NSX Edges. Los objetivos de utilizar balanceadores de carga son por un lado distribuir la carga entre los distintos servidores y por otro mejorar la alta disponibilidad de los servicios si uno de los servidores falla.

Al desplegarse utilizando appliances NSX Edge se pueden desplegar un par de ellos en modo activo-pasivo para asegurar la alta disponibilidad.

Las principales características del balanceador de carga ofrecido en el NSX Edge son:

- Soporte de cualquier aplicación TCP incluyendo servicios como HTTP, HTTPS, FTP…
- Soporte de cualquier aplicación UDP.
- Distintos métodos de balanceo: Round-Robin, least connections, source IP hash y URI.
- Distintos monitores para los servidores: TCP, UDP, HTTP Y HTTPS incluyendo inspección del contenido.
- Persistencia: Por IP de origen, MSRDP, utilizando cookies o ID de sesión SSL.
- Controlo de conexiones: Conexiones máximas o conexiones por segundo.
- Manipulación en capa 7, incluyendo bloqueo de URLs, reescritura de URLs o reescritura del contenido.
-

Como en el caso de los firewalls también permite la integración de balanceadores de otros proveedores.

A la hora de desplegar el NSX Edge para ser utilizado como balanceador existen dos modos:

- One-arm (modo Proxy): El despliegue del NSX Edge se hace directamente en la red lógica donde se tiene que realizar el balanceo.

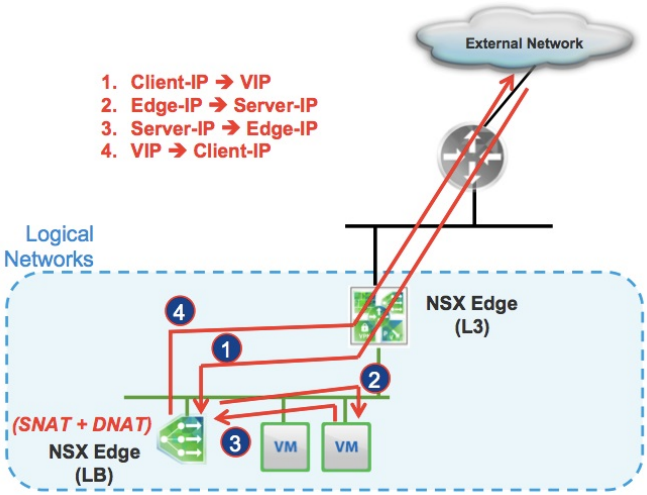

Fig. D.11. Balanceador lógico. Modo One-arm

Este modo es el más simple, permite desplegar el balanceador sin necesidad de modificar el enrutamiento en la infraestructura. Su principal desventaja es que obliga a desplegar más instancias de NSX Edge y que al ser necesario SNAT (para evitar problemas de enrutamiento asimétrico) los servidores de destino no pueden ver la IP de origen original.

- Modo Transparente (en línea): El despliegue del NSX Edge se hace en el camino de la comunicación hacia la granja de servidores:

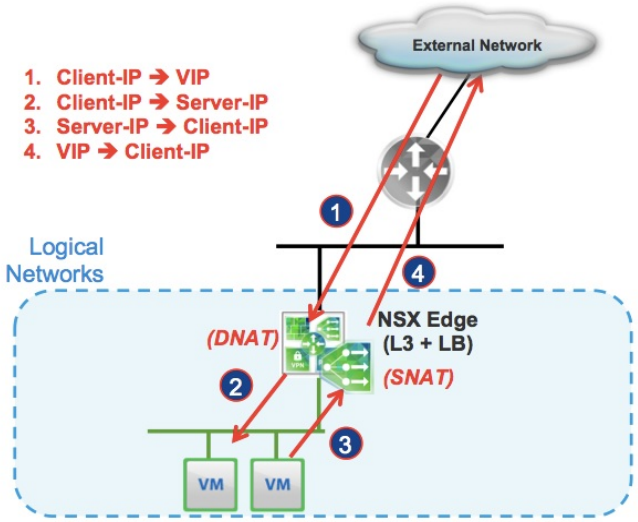

Fig. D.12. Balanceador Lógico. Modo transparente.

Este modo también es bastante simple, y además permite que los servidores vean las IPs originales de los clientes. Como desventajas es menos flexible, ya que el balanceador se convierte en la puerta de enlace para el segmento de red donde están los servidores. Esto implica que no se puede utilizar routing distribuido en estos segmentos, solo centralizado.

El servicio permite gestionar en el mejor de los casos hasta 9 Gbps de tráfico, un millón de conexiones concurrentes y 131.000 conexiones nuevas por segundo.

D.3.5. Servicios de Virtual Private Network (VPN)

Es otra funcionalidad que se puede habilitar en los NSX Edges lo que permite también desplegar la funcionalidad en modo activo-pasivo. Pueden ser de dos tipos: VPNs en capa 2 y VPNs en capa 3.

D.3.5.1. VPNs en capa 2 (L2 VPNs)

Permiten la extender la conectividad en capa 2 entre dos localizaciones separadas.

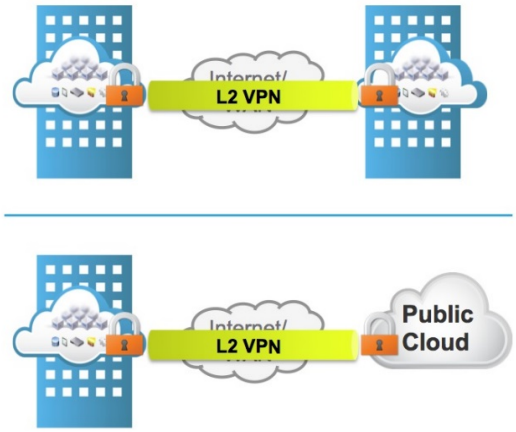

Fig. D.13. VPN L2 en NSX.

Sus principales características son:

- Establecen un túnel SSL entre las dos localizaciones.
- Las redes locales en cada extremo pueden ser de cualquier naturaleza VLANS o VXLANs.
- El NSX Edge en una localización hace de servidor y el de la otra de cliente.
- Es independiente de la red subyacente usada para interconectar las localizaciones que puede ser Internet o enlaces dedicados.
- Se pueden transmitir múltiples VLAN/VXLANs usando trunks.

D.3.5.2. VPNs en capa 3 (L3 VPNs)

Se utiliza para conectar dos redes L3 remotas de forma segura. Pueden ser de dos tipos:

- De acceso remoto. El NSX Edge actúa como servidor. Los clientes se pueden conectar utilizando SSL. (SSL VPN-plus).
- LAN to LAN. Utiliza el protocolo IPSec estándar para conectar con cualquier dispositivo VPN compatible con el protocolo y hacer visibles las redes detrás de cada uno de los peers.

D.3.6. Servicios de conectividad con la red física.

La arquitectura provee de dos tipos de servicio de conectividad de las redes lógicas con las redes físicas. Por un lado, permite dar acceso a redes físicas en capa 3 como pueden ser Internet y por otro, permiten la conectividad de una red lógica contra un segmento en capa 2 físico.

D.3.6.1. Gateway L3

Los servicios de gateway L3 se proveen utilizando el NSX Edge Gateway, como se ha comentado anteriormente usado para proveer a la infraestructura de conectividad Norte-Sur.

Ya que estos appliance se pueden conectar indistintamente a VLANs físicas y VXLANs el equipo proporciona de manera centralizada routing entre las redes lógicas desplegadas en NSX y las redes en la infraestructura física externa. El NSX Edge soporta varios protocolos de enrutamiento dinámicos como OSPF, iBGP o eBGP y también rutas estáticas.

Además, si fuera necesario, el appliance soporta NAT tanto para las IPs de origen como las IPs de destino. Un caso típico sería al interconectar redes privadas (utilizando rangos privados) dentro de la infraestructura con la red pública.

El despliegue de los NSX Edge para esta función se puede hacer en dos modos: modo activo-pasivo y ECMP (Equal Cost MultiPath).

D.3.6.2. Gateway L2 – L2 Bridge

En algunas circunstancias será necesario establecer conexiones en capa 2 entre equipos en el entorno físico y el entorno virtual. Algunos casos típicos son:

- Cuándo hay servidores que por sus características se suelen desplegar directamente sobre servidores físicos, como algunas bases de datos. En

este caso puede ser necesaria la comunicación dentro de la misma VLAN desde el entorno virtual.
- Durante las migraciones de entornos físicos a virtuales.
- Cuando se es necesario utilizar dispositivos de red físicos haciendo de Gateway.
- Al desplegar appliances físicos para ciertas funciones de red (firewalls, IDSs, Balanceadores…).

Para esto, NSX proporciona la funcionalidad de bridge entre VMs en el entorno virtual y equipos en la red física. Esta funcionalidad realiza el mapeo entre la VLAN física y la VXLAN virtual incluso aunque la VLAN no esté presentada en el ESXi dónde está la VM.

Esta funcionalidad (VXLAN-VLAN bridging) se configura a nivel de DLR y se lleva a cabo en el host ESXi en el que se encuentra la DLR Control VM activa. El bridging de los datos se realiza de manera completa en el kernel del host ESXi donde está la DLR control VM.

Algunas de sus características son:

- El mapeo VXLAN-VLAN se realiza siempre en una relación 1 a 1.
- Cada instancia de bridge solo estará activa en un ESXi, el que contenga la DLR Control VM activa.
- La DLR control VM sólo determinará donde se encuentra la instancia de bridging, no participa en ella ya que se lleva a cabo directamente en el kernel del ESXi.
- Se pueden crear múltiples instancias de bridging y se recomienda repartirlas en distintos ESXi para aumentar la escalabilidad.

ANEXO E. ARQUITECTURA FÍSICA PARA LA VIRTUALIZACIÓN CON NSX

E.1. Introducción

El objetivo en la segunda parte del proyecto es definir una arquitectura de red virtual montada sobre la arquitectura NSX de manera alternativa a la arquitectura física propuesta de acuerdo a los requerimientos del cliente.

La infraestructura sobre la que se despliega la solución del cliente, incluyendo la arquitectura de NSX y la red física subyacente la proporcionará el proveedor de servicios en los distintos datacenters y será compartida con otros clientes.

Aunque no es el objetivo de este proyecto definir esta arquitectura, a continuación, se explican a grandes rasgos los requerimientos que tiene que tener la red física sobre la que se despliega NSX y cómo desplegar NSX sobre esta arquitectura de acuerdo a las recomendaciones del fabricante. Para más detalles se recomienda consultar [23].

E.2. Requerimientos de la red física para el despliegue de NSX.

NSX se puede desplegar sobre cualquier topología de red existente. Los únicos requerimientos que son indispensables para que se pueda desplegar la arquitectura es que la red subyacente permita la conectividad IP y que permita el uso de Jumbo-Frames (la encapsulación VXLAN añade cabeceras al paquete ethernet típico y se recomienda al menos utilizar MTUs de 1600 bytes).

En general, los requerimientos para que el despliegue funcione correctamente son: que la red física de transporte IP tenga un ancho de banda alto, que sea tolerante a fallos, que soporte Jumbo-Frames y que la arquitectura soporte calidad de servicio (QoS) para asegurar que los distintos tipos de tráfico (datos, gestión y en algunas ocasiones storage) se prioricen adecuadamente.

E.2.1. Arquitectura de red Modular (clásica).

La arquitectura clásica de DC (llamada Modular) se basa en distintos PODs - *Points of Delivery* (o módulos) en los que hay una capa de acceso y otra de distribución o agregación interconectada utilizando switching en capa 2. La arquitectura puede escalar añadiendo más PODs interconectados utilizando routing en capa 3 por una capa de backbone o core como se muestra en la siguiente figura (Fig. E.1.).

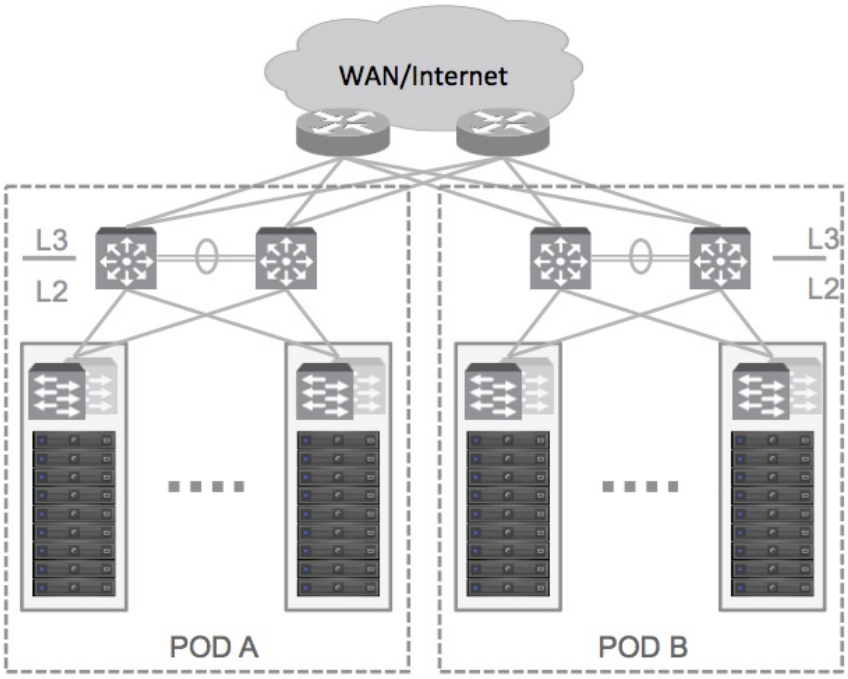

Fig. E.1 Arquitectura clásica de DC basada en acceso, distribución y backbone.

En esta arquitectura las aplicaciones que requieren adyacencia en capa 2 se tienen que desplegar siempre dentro de un mismo POD, ya que cada uno de ellos limita la extensión de las distintas VLANs dentro del DC. Por otro lado, esto implica que los distintos dominios de broadcast L2 están extendidos a lo largo de cada POD con las limitaciones que esto supone: posibilidad de que se formen bucles y redundancia en capa 2 basada en spanning-tree (tiempos de convergencia relativamente altos, enlaces infrautilizados (bloqueados) …). Este diseño es efectivo mientras la mayoría de tráfico se de en sentido norte-sur y no entre los distintos PODs.

La necesidad de incrementar el tráfico este-oeste en el DC y de desplegar aplicaciones sin estar limitadas a los dominios L2 dentro de cada POD ha hecho que el diseño modular evolucione a lo largo de los años hacia un diseño llamado Leaf-Spine.

E.2.2. Arquitectura de red Leaf - Spine

Esta arquitectura se basa en los siguientes componentes:

Capa Leaf: Consiste básicamente en equipos colocados en cada rack (Top of the Rack o TOR) que representan la capa de acceso en la arquitectura. Marcan la frontera entre la red L2 presentada hacia los servidores en el rack y la red L3 en el backbone.

Capa Spine: Es la responsable de interconectar todas las Leafs. Los distintos dispositivos en esta capa no están interconectados ni tienen adyacencias a nivel de routing entre ellos. Normalmente tienen una configuración simple ya que su responsabilidad básica es pasar el tráfico de manera rápida entre las distintas leafs.

Capa border leaf: En algunos escenarios se podrían usar los Spines para interconectar con la red externa pero normalmente el tráfico es retransmitido a varios leafs que se encargan de interconectar con el mundo exterior.

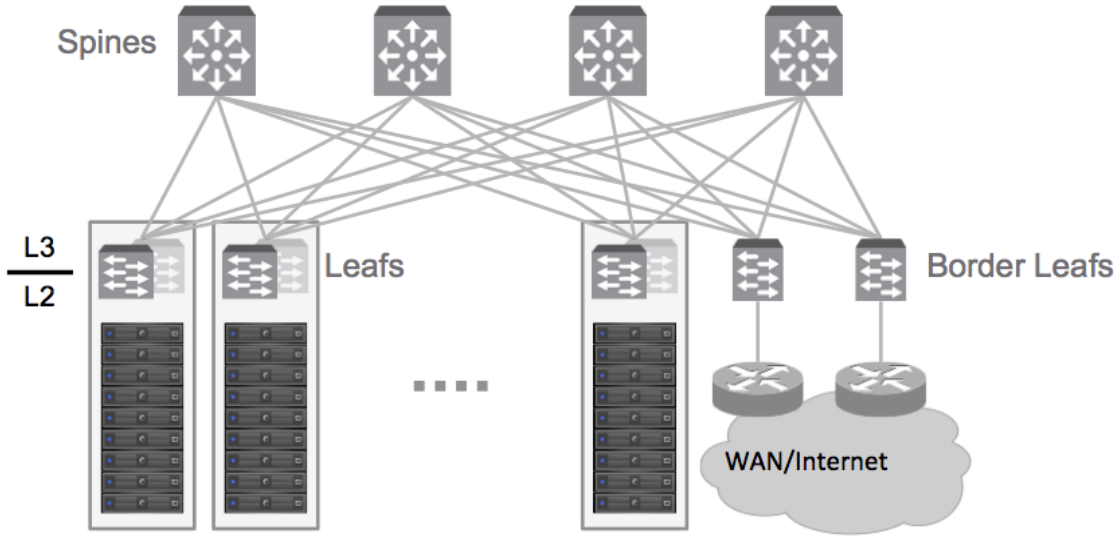

Fig. E.2. Arquitectura Leaf – Spine.

Para interconectar las dos capas en la arquitectura se utiliza enrutamiento dinámico en capa 3. Éste permite la selección del mejor camino en cada momento de acuerdo a los cambios en la red. Gracias al uso de Equal-Cost Multipathing (ECMP) todos los caminos se utilizan de manera simultánea sin crear bucles superando así las limitaciones intrínsecas de utilizar STP. Además, al substituir una red basada en STP por una basada en enrutamiento la red es mucho más estable.

Esta arquitectura es fácil de escalar. Si hay problemas de capacidad en los enlaces entre las dos capas (aumenta la necesidad de tráfico este-oeste) se puede escalar añadiendo un nuevo equipo a la capa Spine conectado contra todos los equipos en la capa leaf. Si hay problemas de capacidad a nivel de puertos en los ToR o switches de acceso se puede añadir un nuevo equipo a la capa de leaf e interconectarlo contra todos los equipos en la capa Spine. Puede haber, por ejemplo, racks donde haya equipos de almacenamiento que requieran más capacidad y por tanto más switches de acceso conectados hacia el backbone.

La arquitectura es tolerante a fallos. Al estar totalmente mallada el fallo de un enlace o un equipo solo supondrá la reducción de la capacidad total.

En general, como habíamos comentado, la topología Leaf-Spine está especialmente indicada para utilizar en arquitecturas en las que es necesario mucho ancho de banda para el tráfico este-oeste. El principal reto para una arquitectura basada en routing como esta es la incapacidad de soportar aplicaciones que requieren estar conectadas en el mismo dominio L2 en cualquier parte del DC. SDN (Software Defined Networks), y la arquitectura NSX en concreto, resuelve este problema gracias a que abstrae la red lógica desplegada de la red física subyacente permitiendo extender los dominios L2 a lo largo del DC sobre esta arquitectura L3 de manera transparente.

E.3. Despliegue de la solución NSX sobre la red física

En este apartado se explica cómo desplegar la virtualización red de NSX sobre una red escalable L3 como es Leaf-Spine, presentada en los apartados anteriores de acuerdo a las recomendaciones del fabricante.

NSX se encargará de proporcionar conectividad L2 utilizando redes lógicas independientes de las características de la red subyacente. Para ello propone una arquitectura en la que se separen y agrupen los distintos ESXi hosts proporcionando funciones específicas como los servidores o máquinas virtuales, los servicios de gestión y los servicios de red.

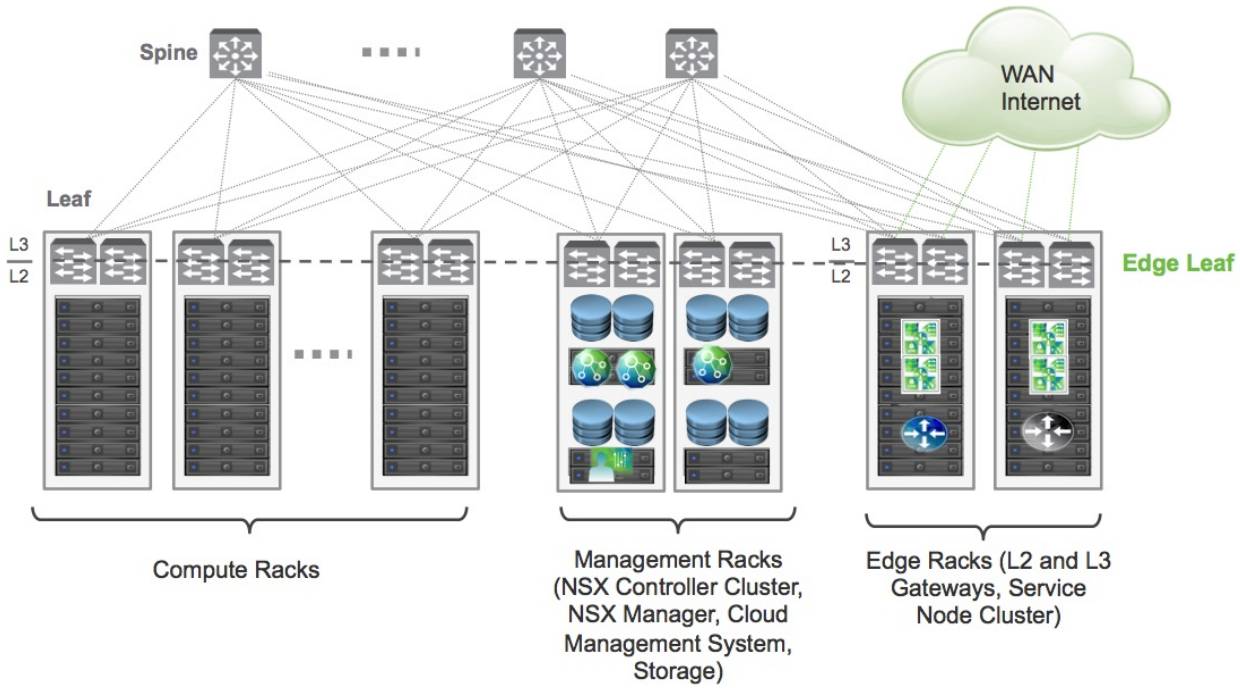

Fig. E.2. NSX sobre arquitectura Leaf-Spine: Compute racks, Management racks y Edge racks.

Compute racks: Contienen los ESXi donde se desplegarán las máquinas virtuales de los clientes:

- No es necesario extender VLANs hasta las máquinas virtuales.
- Se definirán las VLANs de infraestructura de manera local en cada uno de los racks: VXLAN, Vmotion, Storage y gestión. La comunicación de estas VLANs con otros racks se hará de manera enrutada utilizando la arquitectura leaf-spine.
- El modelo de rack será reproducible para cuando sea necesario escalar la solución añadiendo más.

Edge racks: Es donde se hará la interacción entre la red física y la red virtual montada sobre NSX. Las siguientes funciones se realizarán en estos racks:

- Se dará conectividad entre la red física exterior y la red virtual. En el clúster de ESXis en estos racks se desplegarán los NSX Edge Gateways appliances que harán de Gateway L3 hacia la red física.
- Se dará conectividad contra equipos físicos conectados a VLANs en la red física. En este clúster se desplegarán las instancias que realizan las funciones de L2 bridge entre las VXLANs lógicas y las VLANs físicas.
- Se centralizarán los servicios de red ya sean virtuales o físicos (que se vayan a integrar con la infraestructura NSX). Pueden ser equipos como balanceadores, firewalls, IDS…
- Todos los ESXi en este clúster tendrán visibilidad de las VLANs físicas en el datacenter que tengan que tener conectividad contra la infraestructura NSX: VLANs conectadas contra los routers que dan acceso a Internet/WAN, VLANs que se tengan que integrar con las VXLANs en la infraestructura NSX, VLANs que conecten con localizaciones remotas usando servicios en capa 2 y capa 3.

Management racks: Estos hospedarán los servicios de gestión de la infraestructura: vCenter Server para la solución de virtualización, NSX Manager, NSX Controllers, servicios de almacenamiento sobre IP:

- Los componentes de estos racks no tendrán visibilidad ni direccionamiento de las redes especificas usadas por las máquinas virtuales de los clientes.

Es importante destacar que la separación de funciones en los distintos racks es lógica y no necesariamente física. Por ejemplo, en despliegues pequeños se pueden consolidar las funciones/Clústeres de gestión (Management racks) y las de conectividad (Edge racks) en los mismos racks físicos.